

Block-Structured Supermarket Models*

Quan-Lin Li

School of Economics and Management Sciences
Yanshan University, Qinhuangdao 066004, China

John C.S. Lui

Department of Computer Science & Engineering
The Chinese University of Hong Kong, Shatin, N.T, Hong Kong

March 25, 2014

Abstract

Supermarket models are a class of parallel queueing networks with an adaptive control scheme that play a key role in the study of resource management of, such as, computer networks, manufacturing systems and transportation networks. When the arrival processes are non-Poisson and the service times are non-exponential, analysis of such a supermarket model is always limited, interesting, and challenging.

This paper describes a supermarket model with non-Poisson inputs: Markovian Arrival Processes (MAPs) and with non-exponential service times: Phase-type (PH) distributions, and provides a generalized matrix-analytic method which is first combined with the operator semigroup and the mean-field limit. When discussing such a more general supermarket model, this paper makes some new results and advances as follows: (1) Providing a detailed probability analysis for setting up an infinite-dimensional system of differential vector equations satisfied by the expected fraction vector, where *the invariance of environment factors* is given as an important result. (2) Introducing the phase-type structure to the operator semigroup and to the mean-field limit, and a Lipschitz condition can be obtained by means of a unified matrix-differential algorithm. (3) The matrix-analytic method is used to compute the fixed point which leads to performance computation of this system. Finally, we use some

*The main results of this paper will be published in "Discrete Event Dynamic Systems" 2014. On the other hand, the three appendices are the online supplementary material for this paper published in "Discrete Event Dynamic Systems" 2014

numerical examples to illustrate how the performance measures of this supermarket model depend on the non-Poisson inputs and on the non-exponential service times. Thus the results of this paper give new highlight on understanding influence of non-Poisson inputs and of non-exponential service times on performance measures of more general supermarket models.

Keywords: Randomized load balancing; Supermarket model; Matrix-analytic method; Operator semigroup; Mean-field limit; Markovian arrival processes (MAP); Phase-type (PH) distribution; Invariance of environment factors; Doubly exponential tail; *RG*-factorization.

1 Introduction

Supermarket models are a class of parallel queueing networks with an adaptive control scheme that play a key role in the study of resource management of, such as computer networks (e.g., see the dynamic randomized load balancing), manufacturing systems and transportation networks. Since a simple supermarket model was discussed by Mitzenmacher [23], Vvedenskaya et al [32] and Turner [30] through queueing theory as well as Markov processes, subsequent papers have been published on this theme, among which, see, Vvedenskaya and Suhov [33], Jacquet and Vvedenskaya [8], Jacquet et al [9], Mitzenmacher [24], Graham [5, 6, 7], Mitzenmacher et al [25], Vvedenskaya and Suhov [34], Luczak and Norris [20], Luczak and McDiarmid [18, 19], Bramson et al [1, 2, 3], Li et al [17], Li [13] and Li et al [15]. For the fast Jackson networks (or the supermarket networks), readers may refer to Martin and Suhov [22], Martin [21] and Suhov and Vvedenskaya [29].

The available results of the supermarket models with non-exponential service times are still few in the literature. Important examples include an approximate method of integral equations by Vvedenskaya and Suhov [33], the Erlang service times by Mitzenmacher [24] and Mitzenmacher et al [25], the PH service times by Li et al [17] and Li and Lui [16], and the ansatz-based modularized program for the general service times by Bramson et al [1, 2, 3].

Little work has been done on the analysis of the supermarket models with non-Poisson inputs, which are more difficult and challenging due to the higher complexity of that N arrival processes are superposed. Li and Lui [16] and Li [12] used the superposition of N MAP inputs to study the infinite-dimensional Markov processes of supermarket modeling type. Comparing with the results given in Li and Lui [16] and Li [12], this

paper provides more necessary phase-level probability analysis in setting up the infinite-dimensional system of differential vector equations, which leads some new results and methodologies in the study of block-structured supermarket models. Note that the PH distributions constitute a versatile class of distributions that can approximate arbitrarily closely any probability distribution defined on the nonnegative real line, and the MAPs are a broad class of renewal or non-renewal point processes that can approximate arbitrarily closely any stochastic counting process (e.g., see Neuts [27, 28] and Li [11] for more details), thus the results of this paper are a key advance of those given in Mitzenmacher [23] and Vvedenskaya et al [32] under the Poisson and exponential setting.

The main contributions of this paper are threefold. The first one is to use the MAP inputs and the PH service times to describe a more general supermarket model with non-Poisson inputs and with non-exponential service times. Based on the phase structure, we define the random fraction vector and construct an infinite-dimensional Markov process, which expresses the state of this supermarket model by means of an infinite-dimensional Markov process. Furthermore, we set up an infinite-dimensional system of differential vector equations satisfied by the expected fraction vector through a detailed probability analysis. To that end, we obtain an important result: The invariance of environment factors, which is a key for being able to simplify the differential equations in a vector form. Based on the differential vector equations, we can provide a generalized matrix-analytic method to investigate more general supermarket models with non-Poisson inputs and with non-exponential service times. The second contribution of this paper is to provide phase-structured expression for the operator semigroup with respect to the MAP inputs and to the PH service times, and use the operator semigroup to provide the mean-field limit for the sequence of Markov processes who asymptotically approaches a single trajectory identified by the unique and global solution to the infinite-dimensional system of limiting differential vector equations. To prove the existence and uniqueness of solution through the Picard approximation, we provide a unified computational method for establishing a Lipschitz condition, which is crucial in all the rigor proofs involved. The third contribution of this paper is to provide an effective matrix-analytic method both for computing the fixed point and for analyzing performance measures of this supermarket model. Furthermore, we use some numerical examples to indicate how the performance measures of this supermarket model depend on the non-Poisson MAP inputs and on the non-exponential PH service times. Therefore, the results of this paper gives new highlight on understand-

ing performance analysis and nonlinear Markov processes for more general supermarket models with non-Poisson inputs and non-exponential service times.

The remainder of this paper is organized as follows. In Section 2, we first introduce a new MAP whose transition rates are controlled by the number of servers in the system. Then we describe a more general supermarket model of N identical servers with MAP inputs and PH service times. In Section 3, we define a random fraction vector and construct an infinite-dimensional Markov process, which expresses the state of this supermarket model. In Section 4, we set up an infinite-dimensional system of differential vector equations satisfied by the expected fraction vector through a detailed probability analysis, and establish an important result: The invariance of environment factors. In Section 5, we show that the mean-field limit for the sequence of Markov processes who asymptotically approaches a single trajectory identified by the unique and global solution to the infinite-dimensional system of limiting differential vector equations. To prove the existence and uniqueness of the solution, we provide a unified matrix-differential algorithm for establishing the Lipschitz condition. In Section 6, we first discuss the stability of this supermarket model in terms of a coupling method. Then we provide a generalized matrix-analytic method for computing the fixed point whose doubly exponential solution and phase-structured tail are obtained. Finally, we discuss some useful limits of the fraction vector $\mathbf{u}^{(N)}(t)$ as $N \rightarrow \infty$ and $t \rightarrow +\infty$. In Section 7, we provide two performance measures of this supermarket model, and use some numerical examples to indicate how the performance measures of this system depend on the non-Poisson MAP inputs and on the non-exponential PH service times. Some concluding remarks are given in Section 8. Finally, Appendices A and C are respectively designed for the proofs of Theorems 1 and 3, and Appendix B contains the proof of Theorem 2, where the mean-field limit of the sequence of Markov processes in this supermarket model is given a detailed analysis through the operator semigroup.

2 Supermarket Model Description

In this section, we first introduce a new MAP whose transition rates are controlled by the number of servers in the system. Then we describe a more general supermarket model of N identical servers with MAP inputs and PH service times.

2.1 A new Markovian arrival process

Based on Chapter 5 in Neuts [28], the MAP is a bivariate Markov process $\{(N(t), J(t)) : t \geq 0\}$ with state space $S = \{1, 2, 3, \dots\} \times \{1, 2, \dots, m_A\}$, where $\{N(t) : t \geq 0\}$ is a counting process of arrivals and $\{J(t) : t \geq 0\}$ is a Markov environment process. When $J(t) = i$, if the random environment shall go to state j in the next time, then the counting process $\{N(t) : t \geq 0\}$ is a Poisson process with arrival rate $d_{i,j}$ for $1 \leq i, j \leq m_A$. The matrix D with elements $d_{i,j}$ satisfies $D \geq 0$. The matrix C with elements $c_{i,j}$ has negative diagonal elements and nonnegative off-diagonal elements, and the matrix C is invertible, where $c_{i,j}$ is a state transition rate of the Markov chain $\{J(t) : t \geq 0\}$ from state i to state j for $i \neq j$. The matrix $Q = C + D$ is the infinitesimal generator of an irreducible Markov chain. We assume that $Qe = 0$, where e is a column vector of ones with a suitable size. Hence, we have

$$c_{i,i} = - \left[\sum_{j=1}^{m_A} d_{i,j} + \sum_{j \neq i} c_{i,j} \right].$$

Let

$$\mathbb{C} = \begin{pmatrix} -\sum_{j \neq 1}^{m_A} c_{1,j} & c_{1,2} & \cdots & c_{1,m_A} \\ c_{2,1} & -\sum_{j \neq 2}^{m_A} c_{2,j} & \cdots & c_{2,m_A} \\ \vdots & \vdots & \ddots & \vdots \\ c_{m_A,1} & c_{m_A,2} & \cdots & -\sum_{j \neq m_A}^{m_A} c_{m_A,j} \end{pmatrix},$$

$$C(N) = \mathbb{C} - N \text{diag}(De),$$

$$D(N) = ND,$$

where

$$\text{diag}(De) = \text{diag} \left(\sum_{j=1}^{m_A} d_{1,j}, \sum_{j=1}^{m_A} d_{2,j}, \dots, \sum_{j=1}^{m_A} d_{m_A,j} \right).$$

Then

$$Q(N) = C(N) + D(N) = [\mathbb{C} - N \text{diag}(De)] + ND$$

is obviously the infinitesimal generator of an irreducible Markov chain with m_A states. Thus $(C(N), D(N))$ is the irreducible matrix descriptor of a new MAP of order m_A . Note that the new MAP is non-Poisson and may also be non-renewal, and its arrival rate at each environment state is controlled by the number N of servers in the system.

Note that

$$Q(N)e = [\mathbb{C} - N\text{diag}(De)]e + NDe = 0,$$

the Markov chain $Q(N)$ with m_A states is irreducible and positive recurrent. Let ω_N be the stationary probability vector of the Markov chain $Q(N)$. Then ω_N depends on the number $N \geq 1$, and the stationary arrival rate of the MAP is given by $N\lambda_N = N\omega_N De$.

2.2 Model description

Based on the new MAP, we describe a more general supermarket model of N identical servers with MAP inputs and PH service times as follows:

Non-Poisson inputs: Customers arrive at this system as the MAP of irreducible matrix descriptor $(C(N), D(N))$ of size m_A , whose stationary arrival rate is given by $N\lambda_N = N\omega_N De$.

Non-exponential service times: The service times of each server are i.i.d. and are of phase type with an irreducible representation (α, T) of order m_B , where the row vector α is a probability vector whose j th entry is the probability that a service begins in phase j for $1 \leq j \leq m_B$; T is a matrix of size m_B whose $(i, j)^{\text{th}}$ entry is denoted by $t_{i,j}$ with $t_{i,i} < 0$ for $1 \leq i \leq m_B$, and $t_{i,j} \geq 0$ for $i \neq j$. Let $T^0 = -Te = (t_1^0, t_2^0, \dots, t_{m_B}^0)^T \geq 0$, where “ A^T ” denotes the transpose of matrix (or vector) A . When a PH service time is in phase i , the transition rate from phase i to phase j is $t_{i,j}$, the service completion rate is t_i^0 , and the output rate from phase i is $\mu_i = -t_{i,i}$. At the same time, the mean of the PH service time is given by $1/\mu = -\alpha T^{-1}e$.

Arrival and service disciplines: Each arriving customer chooses $d \geq 1$ servers independently and uniformly at random from the N identical servers, and waits for its service at the server which currently contains the fewest number of customers. If there is a tie, servers with the fewest number of customers will be chosen randomly. All customers in any server will be served in the FCFS manner. Figure 1 gives a physical interpretation for this supermarket model.

Remark 1 *The block-structured supermarket models can have many practical applications to, such as, computer networks and manufacturing system, where it is a key to introduce the PH service times and the MAP inputs to such a practical model, because the PH distributions contain many useful distributions such as exponential, hyper-exponential*

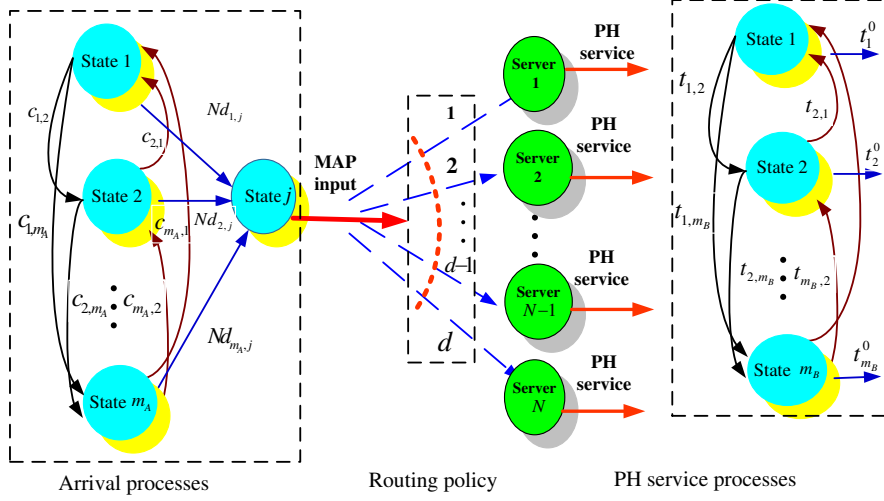


Figure 1: The supermarket model with MAP inputs and PH service times

and Erlang distributions; while the MAPs include, for example, Poisson process, PH-renewal processes, and Markovian modulated Poisson processes (MMPPs). Note that the probability distributions and stochastic point processes have extensively been used in most practical stochastic modeling. On the other hand, in many practical applications, the block-structured supermarket model is an important queueing model to analyze the relation between the system performance and the job routing rule, and it can also help to design reasonable architecture to improve the performance and to balance the load.

3 An Infinite-Dimensional Markov Process

In this section, we first define the random fraction vector of this supermarket model. Then we use the the random fraction vector to construct an infinite-dimensional Markov process, which describes the state of this supermarket model.

For this supermarket model, let $n_{k;i,j}^{(N)}(t)$ be the number of servers with at least k customers (note that the serving customer is also taken into account), and with the MAP be in phase i and the PH service time be in phase j at time $t \geq 0$. Clearly, $0 \leq n_{0;i}^{(N)}(t) \leq N$ and $0 \leq n_{k;i,j}^{(N)}(t) \leq N$ for $k \geq 1$, $1 \leq i \leq m_A$ and $1 \leq j \leq m_B$. Let

$$U_{0;i}^{(N)}(t) = \frac{n_{0;i}^{(N)}(t)}{N}, \quad 1 \leq i \leq m_A,$$

and for $k \geq 1$

$$U_{k;i,j}^{(N)}(t) = \frac{n_{k;i,j}^{(N)}(t)}{N}, \quad 1 \leq i \leq m_A, 1 \leq j \leq m_B.$$

Then $U_{k;i,j}^{(N)}(t)$ is the fraction of servers with at least k customers, and with the MAP be in phase i and the PH service time be in phase j at time t . Using the lexicographic order we write

$$U_0^{(N)}(t) = \left(U_{0;1}^{(N)}(t), U_{0;2}^{(N)}(t), \dots, U_{0;m_A}^{(N)}(t) \right)$$

for $k \geq 1$

$$U_k^{(N)}(t) = \left(U_{k;1,1}^{(N)}(t), U_{k;1,2}^{(N)}(t), \dots, U_{k;1,m_B}^{(N)}(t); \dots; U_{k;m_A,1}^{(N)}(t), U_{k;m_A,2}^{(N)}(t), \dots, U_{k;m_A,m_B}^{(N)}(t) \right),$$

and

$$U^{(N)}(t) = \left(U_0^{(N)}(t), U_1^{(N)}(t), U_2^{(N)}(t), \dots \right). \quad (1)$$

Let $a = (a_1, a_2, a_3, \dots)$ and $b = (b_1, b_2, b_3, \dots)$. We write $a < b$ if $a_k < b_k$ for some $k \geq 1$; $a \leq b$ if $a_k \leq b_k$ for every $k \geq 1$.

For a fixed quaternary array (t, N, i, j) with $t \geq 0, N \in \{1, 2, 3, \dots\}, i \in \{1, 2, \dots, m_A\}$ and $j \in \{1, 2, \dots, m_B\}$, it is easy to see from the stochastic order that $n_{k;i,j}^{(N)}(t) \geq n_{k+1;i,j}^{(N)}(t)$ for $k \geq 1$. This gives

$$U_1^{(N)}(t) \geq U_2^{(N)}(t) \geq U_3^{(N)}(t) \dots \geq 0 \quad (2)$$

and

$$1 = U_0^{(N)}(t)e \geq U_1^{(N)}(t)e \geq U_2^{(N)}(t)e \geq U_3^{(N)}(t)e \geq \dots \geq 0. \quad (3)$$

Note that the state of this supermarket model is described as the random fraction vector $U^{(N)}(t)$ for $t \geq 0$, and $\{U^{(N)}(t), t \geq 0\}$ is a stochastic vector process for each $N = 1, 2, \dots$. Since the arrival process to this supermarket model is the MAP and the service times in each server are of phase type, $\{U^{(N)}(t), t \geq 0\}$ is an infinite-dimensional Markov process whose state space is given by

$$\begin{aligned} \tilde{\Omega}_N = & \left\{ \left(h_0^{(N)}, h_1^{(N)}, h_2^{(N)} \dots \right) : h_0^{(N)} \text{ is a probability vector of size } m_A, \right. \\ & h_1^{(N)} \geq h_2^{(N)} \geq h_3^{(N)} \geq \dots \geq 0, h_k^{(N)} \text{ is a row vector of size } m_A m_B \text{ for } k \geq 1, \\ & 1 = h_0^{(N)}e \geq h_1^{(N)}e \geq h_2^{(N)}e \geq \dots \geq 0, \\ & \left. \text{and } N h_k^{(N)} \text{ is a row vector of nonnegative integers for } k \geq 0 \right\}, \end{aligned} \quad (4)$$

We write

$$u_{0;i}^{(N)}(t) = E \left[U_{0;i}^{(N)}(t) \right]$$

and for $k \geq 1$

$$u_{k;i,j}^{(N)}(t) = E \left[U_{k;i,j}^{(N)}(t) \right].$$

Using the lexicographic order we write

$$u_0^{(N)}(t) = \left(u_{0;1}^{(N)}(t), u_{0;2}^{(N)}(t), \dots, u_{0;m_A}^{(N)}(t) \right)$$

and for $k \geq 1$

$$\begin{aligned} u_k^{(N)}(t) &= \left(u_{k;1,1}^{(N)}(t), u_{k;1,2}^{(N)}(t), \dots, u_{k;1,m_B}^{(N)}(t); \dots; \right. \\ &\quad \left. u_{k;m_A,1}^{(N)}(t), u_{k;m_A,2}^{(N)}(t), \dots, u_{k;m_A,m_B}^{(N)}(t) \right), \\ \mathbf{u}^{(N)}(t) &= \left(u_0^{(N)}(t), u_1^{(N)}(t), u_2^{(N)}(t), \dots \right). \end{aligned}$$

It is easy to see from Equations (2) and (3) that

$$u_1^{(N)}(t) \geq u_2^{(N)}(t) \geq u_3^{(N)}(t) \cdots \geq 0 \quad (5)$$

and

$$1 = u_0^{(N)}(t)e \geq u_1^{(N)}(t)e \geq u_2^{(N)}(t)e \geq \cdots \geq 0. \quad (6)$$

In the remainder of this section, for convenience of readers, it is necessary to explain the structure of this long paper which is outlined as follows. *Part one: The limit of the sequence of Markov processes.* It is seen from (1) and (4) that we need to deal with the limit of the sequence $\{U^{(N)}(t)\}$ of infinite-dimensional Markov processes. This is organized in Appendix B by means of the convergence theorems of operator semigroups, e.g., see Ethier and Kurtz [4] for more details. *Part two: The existence and uniqueness of the solution.* As seen from Theorem 2 and (27), we need to study the two means $E[U^{(N)}(t)]$ and $E[U(t)] = \lim_{N \rightarrow \infty} E[U^{(N)}(t)]$, or $\mathbf{u}^{(N)}(t)$ and $\mathbf{u}(t) = \lim_{N \rightarrow \infty} \mathbf{u}^{(N)}(t)$. To that end, Section 4 sets up the system of differential vector equations satisfied by $\mathbf{u}^{(N)}(t)$, while Section 5 provides a unified matrix-differential algorithm for establishing the Lipschitz condition, which is a key in proving the existence and uniqueness of the solution to the limiting system of differential vector equations satisfied by $\mathbf{u}(t)$ through the Picard approximation. *Part three: Computation of the fixed point and performance analysis.* Section 6 discusses the stability of this supermarket model in terms of a coupling method, and provide an effective matrix-analytic method for computing the fixed point. Section 7 analyzes the performance of this supermarket model by means of some numerical examples.

4 The System of Differential Vector Equations

In this section, we set up an infinite-dimensional system of differential vector equations satisfied by the expected fraction vector through a detailed probability analysis. Specifically, we obtain an important result: The invariance of environment factors, which is a key to rewriting the differential equations as a simple vector form.

To derive the system of differential vector equations, we first discuss an example with the number $k \geq 2$ of customers through the following three steps:

Step one: Analysis of the Arrival Processes

In this supermarket model of N identical servers, we need to determine the change in the expected number of servers with at least k customers over a small time period $[0, dt)$. When the MAP environment process $\{J(t) : t \geq 0\}$ jumps from state l to state i for $1 \leq l, i \leq m_A$ and the PH service environment process $\{I(t) : t \geq 0\}$ sojourns at state j for $1 \leq j \leq m_B$, one arrival occurs in a small time period $[0, dt)$. In this case, the rate that any arriving customer selects d servers with at least $k-1$ customers at random and joins the shortest one with $k-1$ customers, is given by

$$\begin{aligned} & \sum_{l=1}^{m_A} \left[u_{k-1;l,j}^{(N)}(t) d_{l,i} - u_{k;i,j}^{(N)}(t) (d_{i,1}, d_{i,2}, \dots, d_{i,m_A}) e \right] \\ & \times L_{k;l}^{(N)}(u_{k-1}(t), u_k(t)) N dt, \end{aligned} \quad (7)$$

where

$$\begin{aligned} L_{k;l}^{(N)}(u_{k-1}(t), u_k(t)) &= \sum_{m=1}^d C_d^m \left\{ \sum_{j=1}^{m_B} \left[u_{k-1;l,j}^{(N)}(t) - u_{k;l,j}^{(N)}(t) \right] \right\}^{m-1} \left\{ \sum_{j=1}^{m_B} \left[u_{k;l,j}^{(N)}(t) \right] \right\}^{d-m} \\ &+ \sum_{m=1}^{d-1} C_d^m \left\{ \sum_{j=1}^{m_B} \left[u_{k-1;l,j}^{(N)}(t) - u_{k;l,j}^{(N)}(t) \right] \right\}^{m-1} \sum_{\substack{r_1+r_2+\dots+r_{m_A}=d-m \\ \sum_{i \neq l}^{m_A} r_i \geq 1 \\ 0 \leq r_j \leq d-m, 1 \leq j \leq m_A}} \binom{d-m}{r_1, r_2, \dots, r_{m_A}} \\ &\times \prod_{i=1}^{m_A} \left\{ \sum_{j=1}^{m_B} \left[u_{k;i,j}^{(N)}(t) \right] \right\}^{r_i} + \sum_{m=2}^d C_d^m \sum_{m_1=1}^{m-1} \frac{m_1}{m} C_{m_1}^{m_1} \left\{ \sum_{j=1}^{m_B} \left[u_{k-1;l,j}^{(N)}(t) - u_{k;l,j}^{(N)}(t) \right] \right\}^{m_1-1} \end{aligned}$$

| | |
|---|---|
| <p>Each of the d selected servers is at the MAP phase l, and there is at least one server with the shortest queue length $k-1$.</p> <p style="text-align: center;">(Part I)</p> | |
| <p>In the d selected servers, there is only one server with the shortest queue length $k-1$ and with the MAP phase l; there exists at least one server is at the MAP phase $i \neq l$; and all the other $d-1$ selected servers contain no less than k customers.</p> <p style="text-align: center;">(Part II)</p> | <p>In the d selected servers with no less than $k-1$ customers, there is at least one server with the shortest queue length $k-1$, and with the MAP phase l; there is also at least one server with the shortest queue length $k-1$ and with the MAP phase $i \neq l$.</p> <p style="text-align: center;">(Part III)</p> |

Figure 2: A set decomposition of all possible events

$$\begin{aligned}
& \times \sum_{\substack{n_1+n_2+\dots+n_{m_A}=m-m_1 \\ \sum_{i \neq l}^{m_A} n_i \geq 1 \\ 0 \leq n_j \leq m-m_1, 1 \leq j \leq m_A}} \binom{m-m_1}{n_1, n_2, \dots, n_{m_A}} \prod_{i=1}^{m_A} \left\{ \sum_{j=1}^{m_B} \left[u_{k-1;i,j}^{(N)}(t) - u_{k;i,j}^{(N)}(t) \right] \right\}^{n_i} \\
& \times \sum_{\substack{r_1+r_2+\dots+r_{m_A}=d-m \\ 0 \leq r_j \leq d-m, 1 \leq j \leq m_A}} \binom{d-m}{r_1, r_2, \dots, r_{m_A}} \prod_{i=1}^{m_A} \left\{ \sum_{j=1}^{m_B} \left[u_{k;i,j}^{(N)}(t) \right] \right\}^{r_i}. \tag{8}
\end{aligned}$$

Note that $\left[u_{k-1;l,j}^{(N)}(t) d_{l,i} - u_{k;i,j}^{(N)}(t) (d_{i,1}, d_{i,2}, \dots, d_{i,m_A}) e \right]$ is the rate that any arriving customer joins one server with the shortest queue length $k-1$, where the MAP goes to phase i from phase l , and the PH service time is in phase j .

Now, we provide a detailed interpretation for how to derive (8) through a set decomposition of all possible events given in Figure 2, where each of the d selected servers has at least $k-1$ customers, the MAP arrival environment is in phase i or l , and the PH service environment is in phase j . Hence, the probability that any arriving customer selects d servers with at least $k-1$ customers at random and joins a server with the shortest queue length $k-1$ and with the MAP phase i or l is determined by means of Figure 2 through the following three parts:

Part I: The probability that any arriving customer joins a server with the shortest queue length $k-1$ and with the MAP phase l , and the queue lengths of the other selected

$d - 1$ servers are not shorter than $k - 1$, is given by

$$\sum_{m=1}^d C_d^m \left\{ \sum_{j=1}^{m_B} [u_{k-1;l,j}^{(N)}(t) - u_{k;l,j}^{(N)}(t)] \right\}^{m-1} \left\{ \sum_{j=1}^{m_B} [u_{k;l,j}^{(N)}(t)] \right\}^{d-m},$$

where $C_d^m = d!/[m!(d-m)!]$ is a binomial coefficient, and

$$\left\{ \sum_{j=1}^{m_B} [u_{k-1;l,j}^{(N)}(t) - u_{k;l,j}^{(N)}(t)] \right\}^{m-1}$$

is the probability that any arriving customer who can only choose one server makes $m - 1$ independent selections during the $m - 1$ servers with the queue length $k - 1$ and with the MAP phase l at time t ; while $\left\{ \sum_{j=1}^{m_B} [u_{k;l,j}^{(N)}(t)] \right\}^{d-m}$ is the probability that there are $d - m$ servers whose queue lengths are not shorter than k and with the MAP phase l .

Part II: The probability that any arriving customer joins a server with the shortest queue length $k - 1$ and with the MAP phase l ; and the queue lengths of the other selected $d - 1$ servers are not shorter than $k - 1$, and there exist at least one server with no less than k customers and with the MAP phase $i \neq l$, is given by

$$\begin{aligned} & \sum_{m=1}^{d-1} C_d^m \left\{ \sum_{j=1}^{m_B} [u_{k-1;l,j}^{(N)}(t) - u_{k;l,j}^{(N)}(t)] \right\}^{m-1} \\ & \times \sum_{\substack{r_1+r_2+\dots+r_{m_A}=d-m \\ \sum_{i \neq l}^{m_A} r_i \geq 1 \\ 0 \leq r_j \leq d-m, 1 \leq j \leq m_A}} \binom{d-m}{r_1, r_2, \dots, r_{m_A}} \prod_{i=1}^{m_A} \left\{ \sum_{j=1}^{m_B} [u_{k;i,j}^{(N)}(t)] \right\}^{r_i}, \end{aligned}$$

where when $r_1 + r_2 + \dots + r_{m_A} = n$, $\binom{n}{r_1, r_2, \dots, r_{m_A}} = \frac{n!}{\prod_{i=1}^{m_A} r_i!}$ is a multinomial coefficient.

Part III: If there are m selected servers with the shortest queue length $k - 1$ where there are m_1 servers with the MAP phase l and $m - m_1$ servers with the MAP phases $i \neq l$, then the probability that any arriving customer joins a server with the shortest queue length $k - 1$ and with the MAP phase l is equal to m_1/m . In this case, the probability that any arriving customer joins a server with the shortest queue length $k - 1$ and with the MAP phase l , the queue lengths of the other selected $d - 1$ servers are not shorter than $k - 1$,

is given by

$$\begin{aligned}
& \sum_{m=2}^d C_d^m \sum_{m_1=1}^{m-1} \frac{m_1}{m} C_m^{m_1} \left\{ \sum_{j=1}^{m_B} \left[u_{k-1;l,j}^{(N)}(t) - u_{k;l,j}^{(N)}(t) \right] \right\}^{m_1-1} \\
& \times \sum_{\substack{n_1+n_2+\dots+n_{m_A}=m-m_1 \\ \sum_{i \neq l}^{m_A} n_i \geq 1 \\ 0 \leq n_j \leq m-m_1, 1 \leq j \leq m_A}} \binom{m-m_1}{n_1, n_2, \dots, n_{m_A}} \prod_{i=1}^{m_A} \left\{ \sum_{j=1}^{m_B} \left[u_{k-1;i,j}^{(N)}(t) - u_{k;i,j}^{(N)}(t) \right] \right\}^{n_i} \\
& \times \sum_{\substack{r_1+r_2+\dots+r_{m_A}=d-m \\ 0 \leq r_j \leq d-m, 1 \leq j \leq m_A}} \binom{d-m}{r_1, r_2, \dots, r_{m_A}} \prod_{i=1}^{m_A} \left\{ \sum_{j=1}^{m_B} \left[u_{k;i,j}^{(N)}(t) \right] \right\}^{r_i}.
\end{aligned}$$

Using the above three parts, (7) and (8) can be obtained immediately.

For any two matrices $A = (a_{i,j})$ and $B = (b_{i,j})$, their Kronecker product is defined as $A \otimes B = (a_{i,j} B)$, and their Kronecker sum is given by $A \oplus B = A \otimes I + I \otimes B$.

The following theorem gives an important result, called *the invariance of environment factors*, which will play an important role in setting up the infinite-dimensional system of differential vector equations. This enables us to apply the matrix-analytic method to the study of more general supermarket models with non-Poisson inputs and non-exponential service times.

Theorem 1

$$\begin{aligned}
L_{1;l}^{(N)} \left(u_0^{(N)}(t) \otimes \alpha, u_1^{(N)}(t) \right) &= \sum_{m=1}^d C_d^m \left[\sum_{l=1}^{m_A} \sum_{j=1}^{m_B} \left(u_{0;l}^{(N)}(t) \alpha_j - u_{1;l,j}^{(N)}(t) \right) \right]^{m-1} \\
&\times \left[\sum_{l=1}^{m_A} \sum_{j=1}^{m_B} u_{1;l,j}^{(N)}(t) \right]^{d-m}
\end{aligned} \tag{9}$$

and for $k \geq 2$

$$\begin{aligned}
L_{k;l}^{(N)} \left(u_{k-1}^{(N)}(t), u_k^{(N)}(t) \right) &= \sum_{m=1}^d C_d^m \left[\sum_{l=1}^{m_A} \sum_{j=1}^{m_B} \left(u_{k-1;l,j}^{(N)}(t) - u_{k;l,j}^{(N)}(t) \right) \right]^{m-1} \\
&\times \left[\sum_{l=1}^{m_A} \sum_{j=1}^{m_B} u_{k;l,j}^{(N)}(t) \right]^{d-m}.
\end{aligned} \tag{10}$$

Thus $L_{1;l}^{(N)} \left(u_0^{(N)}(t) \otimes \alpha, u_1^{(N)}(t) \right)$ and $L_{k;l}^{(N)} \left(u_{k-1}^{(N)}(t), u_k^{(N)}(t) \right)$ for $k \geq 2$ are independent of the MAP phase $l \in \{1, 2, \dots, m_A\}$. In this case, we have

$$L_{1;l}^{(N)} \left(u_0^{(N)}(t) \otimes \alpha, u_1^{(N)}(t) \right) \stackrel{\text{def}}{=} L_1^{(N)} \left(u_0^{(N)}(t) \otimes \alpha, u_1^{(N)}(t) \right) \tag{11}$$

and for $k \geq 2$

$$L_{k;l}^{(N)} \left(u_{k-1}^{(N)}(t), u_k^{(N)}(t) \right) \stackrel{\text{def}}{=} L_k^{(N)} \left(u_{k-1}^{(N)}(t), u_k^{(N)}(t) \right). \quad (12)$$

Proof: See Appendix A. ■

It is seen from the invariance of environment factors in Theorem 1 that Equation (7) is rewritten as, in a vector form,

$$\begin{aligned} & \left\{ u_{k-1}^{(N)}(t) (D \otimes I) - u_k^{(N)}(t) [\text{diag}(De) \otimes I] \right\} \\ & \times L_k^{(N)} \left(u_{k-1}^{(N)}(t), u_k^{(N)}(t) \right) N dt. \end{aligned} \quad (13)$$

Note that $L_1^{(N)} \left(u_0^{(N)}(t) \otimes \alpha, u_1^{(N)}(t) \right)$ and $L_k^{(N)} \left(u_{k-1}^{(N)}(t), u_k^{(N)}(t) \right)$ are scale for $k \geq 2$.

Step two: Analysis of the Environment State Transitions in the MAP

When there are at least k customers in the server, the rate that the MAP environment process jumps from state l to state i with rate $c_{l,j}$, and no arrival of the MAP occurs during a small time period $[0, dt)$, is given by

$$\left[\sum_{l=1}^{m_A} u_{k;l,j}^{(N)}(t) c_{l,i} + u_{k,i,j}^{(N)}(t) (d_{i,1}, d_{i,2}, \dots, d_{i,m_A}) e \right] N dt.$$

This gives, in a vector form,

$$u_k^{(N)}(t) ([C + \text{diag}(De)] \otimes I) N dt. \quad (14)$$

Step three: Analysis of the Service Processes

To analyze the PH service process, we need to consider the following two cases:

Case one: One service completion occurs with rate t_l^0 during a small time period $[0, dt)$. In this case, when there are at least $k+1$ customers in the server, the rate that a customer is completed its service with entering PH phase j and the MAP is in phase i is given by

$$\left[u_{k+1;i,1}^{(N)}(t) t_1^0 \alpha_j + u_{k+1;i,2}^{(N)}(t) t_2^0 \alpha_j + \dots + u_{k+1;i,m_B}^{(N)}(t) t_{m_B}^0 \alpha_j \right] N dt.$$

Case two: No service completion occurs during a small time period $[0, dt)$, but the MAP is in phase i and the PH service environment process goes to phase j . Thus, when there are at least k customers in the server, the rate of this case is given by

$$\left[u_{k;i,1}^{(N)}(t) t_{1,j} + u_{k;i,2}^{(N)}(t) t_{2,j} + u_{k;i,3}^{(N)}(t) t_{3,j} + \dots + u_{k;i,m_B}^{(N)}(t) t_{m_B,j} \right] N dt.$$

Thus, for the PH service process, we obtain that in a vector form,

$$\left[u_k^{(N)}(t) (I \otimes T) + u_{k+1}^{(N)}(t) (I \otimes T^0 \alpha) \right] N dt \quad (15)$$

Let

$$n_k^{(N)}(t) = \left(n_{k;1,1}^{(N)}(t), n_{k;1,2}^{(N)}(t), \dots, n_{k;1,m_B}^{(N)}(t); \dots; n_{k;m_A,1}^{(N)}(t), n_{k;m_A,2}^{(N)}(t), \dots, n_{k;m_A,m_B}^{(N)}(t) \right).$$

Then it follows from Equation (13) to (15) that

$$\begin{aligned} dE \left[n_k^{(N)}(t) \right] = & \left\{ \left\{ u_{k-1}^{(N)}(t) (D \otimes I) - u_k^{(N)}(t) [\text{diag}(De) \otimes I] \right\} L_k^{(N)} \left(u_{k-1}^{(N)}(t), u_k^{(N)}(t) \right) \right. \\ & \left. + u_k^{(N)}(t) \{ [C + \text{diag}(De)] \oplus T \} + u_{k+1}^{(N)}(t) (I \otimes T^0 \alpha) \right\} N dt. \end{aligned}$$

Since $E \left[n_k^{(N)}(t)/N \right] = u_k^{(N)}(t)$ and $A \otimes I + I \otimes B = A \oplus B$, we obtain

$$\begin{aligned} \frac{du_k^{(N)}(t)}{dt} = & \left\{ u_{k-1}^{(N)}(t) (D \otimes I) - u_k^{(N)}(t) [\text{diag}(De) \otimes I] \right\} L_k^{(N)} \left(u_{k-1}^{(N)}(t), u_k^{(N)}(t) \right) \\ & + u_k^{(N)}(t) \{ [C + \text{diag}(De)] \oplus T \} + u_{k+1}^{(N)}(t) (I \otimes T^0 \alpha). \end{aligned} \quad (16)$$

Using a similar analysis to Equation (16), we obtain an infinite-dimensional system of differential vector equations satisfied by the expected fraction vector $\mathbf{u}^{(N)}(t)$ as follows:

$$\begin{aligned} \frac{du_1^{(N)}(t)}{dt} = & \left\{ \left[u_0^{(N)}(t) \otimes \alpha \right] (D \otimes I) - u_1^{(N)}(t) [\text{diag}(De) \otimes I] \right\} L_1^{(N)} \left(u_0^{(N)}(t) \otimes \alpha, u_1^{(N)}(t) \right) \\ & + u_1^{(N)}(t) \{ [C + \text{diag}(De)] \oplus T \} + u_2^{(N)}(t) (I \otimes T^0 \alpha), \end{aligned} \quad (17)$$

and for $k \geq 2$

$$\begin{aligned} \frac{du_k^{(N)}(t)}{dt} = & \left\{ u_{k-1}^{(N)}(t) (D \otimes I) - u_k^{(N)}(t) [\text{diag}(De) \otimes I] \right\} L_k^{(N)} \left(u_{k-1}^{(N)}(t), u_k^{(N)}(t) \right) \\ & + u_k^{(N)}(t) \{ [C + \text{diag}(De)] \oplus T \} + u_{k+1}^{(N)}(t) (I \otimes T^0 \alpha), \end{aligned} \quad (18)$$

with the boundary condition

$$\frac{du_0^{(N)}(t)}{dt} = u_0^{(N)}(t) (C + D), \quad (19)$$

$$u_0^{(N)}(t)e = 1; \quad (20)$$

and with the initial condition

$$u_k^{(N)}(0) = g_k, \quad k \geq 1, \quad (21)$$

where

$$g_1 \geq g_2 \geq g_3 \geq \dots \geq 0$$

and

$$1 = g_0 e \geq g_1 e \geq g_2 e \geq \dots \geq 0.$$

Remark 2 *It is necessary to explain some probability setting for the invariance of environment factors. It follows from Theorem 1 that*

$$L_1^{(N)} \left(u_0^{(N)}(t) \otimes \alpha, u_1^{(N)}(t) \right) = \frac{\left[u_0^{(N)}(t) e \right]^d - \left[u_1^{(N)}(t) e \right]^d}{u_0^{(N)}(t) e - u_1^{(N)}(t) e}$$

and for $k \geq 2$

$$L_k^{(N)} \left(u_{k-1}^{(N)}(t), u_k^{(N)}(t) \right) = \frac{\left[u_{k-1}^{(N)}(t) e \right]^d - \left[u_k^{(N)}(t) e \right]^d}{u_{k-1}^{(N)}(t) e - u_k^{(N)}(t) e}.$$

Note that the two expressions will be useful in our later study, for example, establishing the Lipschitz condition, and computing the fixed point. Specifically, for $d = 1$ we have

$$L_1^{(N)} \left(u_0^{(N)}(t) \otimes \alpha, u_1^{(N)}(t) \right) = 1$$

and for $k \geq 2$

$$L_k^{(N)} \left(u_{k-1}^{(N)}(t), u_k^{(N)}(t) \right) = 1.$$

For $d = 2$ we have

$$L_1^{(N)} \left(u_0^{(N)}(t) \otimes \alpha, u_1^{(N)}(t) \right) = u_0^{(N)}(t) e + u_1^{(N)}(t) e > 1$$

and for $k \geq 2$

$$L_k^{(N)} \left(u_{k-1}^{(N)}(t), u_k^{(N)}(t) \right) = u_{k-1}^{(N)}(t) e + u_k^{(N)}(t) e.$$

This shows that $\left(L_1^{(N)} \left(u_0^{(N)}(t) \otimes \alpha, u_1^{(N)}(t) \right), L_2^{(N)} \left(u_1^{(N)}(t), u_2^{(N)}(t) \right), \dots \right)$ is not a probability vector.

5 The Lipschitz Condition

In this section, we show that the mean-field limit of the sequence of Markov processes asymptotically approaches a single trajectory identified by the unique and global solution to the infinite-dimensional system of limiting differential vector equations. To that end,

we provide a unified matrix-differential algorithm for establishing the Lipschitz condition, which is a key in proving the existence and uniqueness of the solution by means of the Picard approximation according to the basic results of the Banach space.

Let $\mathbf{T}_N(t)$ be the operator semigroup of the Markov process $\{\mathbf{U}^{(N)}(t), t \geq 0\}$. If $f : \Omega_N \rightarrow \mathbf{C}^1$, where $\Omega_N = \{\mathbf{g} \in \tilde{\Omega}_N : \mathbf{g}e < +\infty\}$, then for $\mathbf{g} \in \Omega_N$ and $t \geq 0$

$$\mathbf{T}_N(t)f(\mathbf{g}) = E[f(\mathbf{U}_N(t) \mid \mathbf{U}_N(0) = \mathbf{g})].$$

We denote by \mathbf{A}_N the generating operator of the operator semigroup $\mathbf{T}_N(t)$, it is easy to see that $\mathbf{T}_N(t) = \exp\{\mathbf{A}_N t\}$ for $t \geq 0$. In **Appendix B**, we will provide a detailed analysis for the limiting behavior of the sequence $\{(\mathbf{U}^{(N)}(t), t \geq 0\}$ of Markov processes for $N = 1, 2, 3, \dots$, where two formal limits for the sequence $\{\mathbf{A}_N\}$ of generating operators and for the sequence $\{\mathbf{T}_N(t)\}$ of operator semigroups are expressed as $\mathbf{A} = \lim_{N \rightarrow \infty} \mathbf{A}_N$ and $\mathbf{T}(t) = \lim_{N \rightarrow \infty} \mathbf{T}_N(t)$ for $t \geq 0$, respectively.

We write

$$L_1(u_0(t) \otimes \alpha, u_1(t)) = \sum_{m=1}^d C_d^m \left[\sum_{l=1}^{m_A} \sum_{j=1}^{m_B} (u_{0,l}(t) \alpha_j - u_{1;l,j}(t)) \right]^{m-1} \left[\sum_{l=1}^{m_A} \sum_{j=1}^{m_B} u_{1;l,j}(t) \right]^{d-m},$$

for $k \geq 2$

$$\begin{aligned} L_k(u_{k-1}(t), u_k(t)) &= \sum_{m=1}^d C_d^m \left[\sum_{l=1}^{m_A} \sum_{j=1}^{m_B} (u_{k-1;l,j}(t) - u_{k;l,j}(t)) \right]^{m-1} \\ &\quad \times \left[\sum_{l=1}^{m_A} \sum_{j=1}^{m_B} u_{k;l,j}(t) \right]^{d-m}. \end{aligned}$$

Let $\mathbf{u}(t) = \lim_{N \rightarrow \infty} \mathbf{u}^{(N)}(t)$ where $u_k(t) = \lim_{N \rightarrow \infty} u_k^{(N)}(t)$ for $k \geq 0$ and $t \geq 0$. Based on the limiting operator semigroup $\mathbf{T}(t)$ or the limiting generating operator \mathbf{A} , as $N \rightarrow \infty$ it follows from Equations (17) to (21) that $\mathbf{u}(t)$ is a solution to the system of differential vector equations as follows:

$$\begin{aligned} \frac{du_1(t)}{dt} &= \{[u_0(t) \otimes \alpha] (D \otimes I) - u_1(t) [\text{diag}(De) \otimes I]\} L_1(u_0(t) \otimes \alpha, u_1(t)) \\ &\quad + u_1(t) \{[C + \text{diag}(De)] \oplus T\} + u_2(t) (I \otimes T^0 \alpha), \end{aligned} \quad (22)$$

and for $k \geq 2$

$$\begin{aligned} \frac{du_k(t)}{dt} &= \{u_{k-1}(t) (D \otimes I) - u_k(t) [\text{diag}(De) \otimes I]\} L_k(u_{k-1}(t), u_k(t)) \\ &\quad + u_k(t) \{[C + \text{diag}(De)] \oplus T\} + u_{k+1}(t) (I \otimes T^0 \alpha), \end{aligned} \quad (23)$$

with the boundary condition

$$u_0^{(N)}(t) = u_0^{(N)}(0) \exp\{(C + D)t\}, \quad (24)$$

$$u_0^{(N)}(t)e = 1, \quad (25)$$

and with initial condition

$$u_k(0) = g_k, \quad k \geq 0. \quad (26)$$

Based on the solution $\mathbf{u}(t, \mathbf{g})$ to the system of differential vector equations (22) to (26), we define a mapping: $\mathbf{g} \rightarrow \mathbf{u}(t, \mathbf{g})$. Note that the operator semigroup $\mathbf{T}(t)$ acts in the space L , where $L = C(\tilde{\Omega})$ is the Banach space of continuous functions $f : \tilde{\Omega} \rightarrow \mathbf{R}$ with uniform metric $\|f\| = \max_{u \in \tilde{\Omega}} |f(u)|$, and

$$\tilde{\Omega} = \{\mathbf{u} : u_1 \geq u_2 \geq u_3 \geq \dots \geq 0; \quad 1 = u_0^{(N)}e \geq u_1^{(N)}e \geq u_2^{(N)}e \geq \dots \geq 0\}$$

for the vector $\mathbf{u} = (u_0, u_1, u_2, \dots)$ with u_0 be a probability vector of size m_A and the size of the row vector u_k be $m_A m_B$ for $k \geq 1$. If $f \in L$ and $\mathbf{g} \in \tilde{\Omega}$, then

$$\mathbf{T}(t)f(\mathbf{g}) = f(\mathbf{u}(t, \mathbf{g})).$$

The following theorem uses the operator semigroup to provide the mean-field limit in this supermarket model. Note that the mean-field limit shows that there always exists the limiting process $\{U(t), t \geq 0\}$ of the sequence $\{U^{(N)}(t), t \geq 0\}$ of Markov processes, and also indicates the asymptotic independence of the block-structured queueing processes in this supermarket model.

Theorem 2 *For any continuous function $f : \Omega \rightarrow \mathbf{R}$ and $t > 0$,*

$$\lim_{N \rightarrow \infty} \sup_{\mathbf{g} \in \Omega} |\mathbf{T}_N(t)f(\mathbf{g}) - f(\mathbf{u}(t; \mathbf{g}))| = 0,$$

and the convergence is uniform in t with any bounded interval.

Proof: See Appendix B. ■

Finally, we provide some interpretation on Theorem 2. If $\lim_{N \rightarrow \infty} U^{(N)}(0) = \mathbf{u}(0) = \mathbf{g} \in \Omega$ in probability, then Theorem 2 shows that $U(t) = \lim_{N \rightarrow \infty} U^{(N)}(t)$ is concentrated on the trajectory $\Gamma_{\mathbf{g}} = \{\mathbf{u}(t, \mathbf{g}) : t \geq 0\}$. This indicates the functional strong law of large numbers for the time evolution of the fraction of each state of this supermarket model,

thus the sequence $\{U^{(N)}(t), t \geq 0\}$ of Markov processes converges weakly to the expected fraction vector $\mathbf{u}(t, \mathbf{g})$ as $N \rightarrow \infty$, that is, for any $T > 0$

$$\lim_{N \rightarrow \infty} \sup_{0 \leq s \leq T} \|U^{(N)}(s) - \mathbf{u}(s, \mathbf{g})\| = 0 \quad \text{in probability.} \quad (27)$$

In the remainder of this section, we provide a unified matrix-differential algorithm for establishing a Lipschitz condition for the expected fraction vector $f : \mathbf{R}_+^\infty \rightarrow \mathbf{C}^1(\mathbf{R}_+^\infty)$. The Lipschitz condition is a key for proving the existence and uniqueness of solution to the infinite-dimensional system of limiting differential vector equations (22) to (26). On the other hand, the proof of the existence and uniqueness of solution is standard by means of the Picard approximation according to the basic results of the Banach space. Readers may refer to Li, Dai, Lui and Wang [15] for more details.

To provide the Lipschitz condition, we need to use the derivative of the infinite-dimensional vector $G : \mathbf{R}_+^\infty \rightarrow \mathbf{C}^1(\mathbf{R}_+^\infty)$. Thus we first provide some definitions and preliminaries for such derivatives as follows.

For the infinite-dimensional vector $G : \mathbf{R}_+^\infty \rightarrow \mathbf{C}^1(\mathbf{R}_+^\infty)$, we write $x = (x_1, x_2, x_3, \dots)$ and $G(x) = (G_1(x), G_2(x), G_3(x), \dots)$, where x_k and $G_k(x)$ are scalar for $k \geq 1$. Then the matrix of partial derivatives of the infinite-dimensional vector $G(x)$ is defined as

$$\mathcal{D}G(x) = \frac{\partial G(x)}{\partial x} = \begin{pmatrix} \frac{\partial G_1(x)}{\partial x_1} & \frac{\partial G_2(x)}{\partial x_1} & \frac{\partial G_3(x)}{\partial x_1} & \dots \\ \frac{\partial G_1(x)}{\partial x_2} & \frac{\partial G_2(x)}{\partial x_2} & \frac{\partial G_3(x)}{\partial x_2} & \dots \\ \frac{\partial G_1(x)}{\partial x_3} & \frac{\partial G_2(x)}{\partial x_3} & \frac{\partial G_3(x)}{\partial x_3} & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}, \quad (28)$$

if each of the partial derivatives exists.

For the infinite-dimensional vector $G : \mathbf{R}_+^\infty \rightarrow \mathbf{C}^1(\mathbf{R}_+^\infty)$, if there exists a linear operator $A : \mathbf{R}_+^\infty \rightarrow \mathbf{C}^1(\mathbf{R}_+^\infty)$ such that for any vector $h \in \mathbf{R}_+^\infty$ and a scalar $t \in \mathbf{R}$

$$\lim_{t \rightarrow 0} \frac{\|G(x + th) - G(x) - thA\|}{t} = 0,$$

then the function $G(x)$ is called to be Gateaux differentiable at $x \in \mathbf{R}_+^\infty$. In this case, we write the Gateaux derivative $A = \mathcal{D}G(x) = \frac{\partial G(x)}{\partial x}$.

Let $\mathbf{t} = (t_1, t_2, t_3, \dots)$ with $0 \leq t_k \leq 1$ for $k \geq 1$. Then we write

$$\mathcal{D}G(x + \mathbf{t} \oslash (y - x)) = \begin{pmatrix} \frac{\partial G_1(x + t_1(y - x))}{\partial x_1} & \frac{\partial G_2(x + t_2(y - x))}{\partial x_1} & \frac{\partial G_3(x + t_3(y - x))}{\partial x_1} & \dots \\ \frac{\partial G_1(x + t_1(y - x))}{\partial x_2} & \frac{\partial G_2(x + t_2(y - x))}{\partial x_2} & \frac{\partial G_3(x + t_3(y - x))}{\partial x_2} & \dots \\ \frac{\partial G_1(x + t_1(y - x))}{\partial x_3} & \frac{\partial G_2(x + t_2(y - x))}{\partial x_3} & \frac{\partial G_3(x + t_3(y - x))}{\partial x_3} & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

If the infinite-dimensional vector $G : \mathbf{R}_+^\infty \rightarrow \mathbf{C}^1(\mathbf{R}_+^\infty)$ is Gateaux differentiable, then there exists a vector $\mathbf{t} = (t_1, t_2, t_3, \dots)$ with $0 \leq t_k \leq 1$ for $k \geq 1$ such that

$$G(y) - G(x) = (y - x) \mathcal{D}G(x + \mathbf{t} \oslash (y - x)). \quad (29)$$

Furthermore, we have

$$\|G(y) - G(x)\| \leq \sup_{0 \leq t \leq 1} \|\mathcal{D}G(x + t(y - x))\| \|y - x\|. \quad (30)$$

For convenience of description, Equations (22) to (26) are rewritten as an initial value problem as follows:

$$\begin{aligned} \frac{d}{dt} u_1 &= \{(u_0 \otimes \alpha)(D \otimes I) - u_1 [\text{diag}(De) \otimes I]\} L_1(u_0 \otimes \alpha, u_1) \\ &\quad + u_1 \{[C + \text{diag}(De)] \oplus T\} + u_2 (I \otimes T^0 \alpha) \end{aligned} \quad (31)$$

and for $k \geq 2$,

$$\begin{aligned} \frac{d}{dt} u_k &= \{u_{k-1} (D \otimes I) - u_k [\text{diag}(De) \otimes I]\} L_k(u_{k-1}, u_k) \\ &\quad + u_k \{[C + \text{diag}(De)] \oplus T\} + u_{k+1} (I \otimes T^0 \alpha), \end{aligned} \quad (32)$$

with the initial condition

$$u_k(0) = g_k, \quad k \geq 0, \quad (33)$$

where for $t \geq 0$

$$u_0(t) = u_0(0) \exp\{(C + D)t\}$$

and

$$u_0(t) e = 1.$$

Let $x = (x_1, x_2, x_3, \dots) = (u_1, u_2, u_3, \dots)$ and $F(x) = (F_1(x), F_2(x), F_3(x), \dots)$, where

$$\begin{aligned} F_1(x) &= \{(u_0 \otimes \alpha)(D \otimes I) - x_1 [\text{diag}(De) \otimes I]\} L_1(u_0 \otimes \alpha, x_1) \\ &\quad + x_1 \{[C + \text{diag}(De)] \oplus T\} + x_2 (I \otimes T^0 \alpha) \end{aligned} \quad (34)$$

and for $k \geq 2$

$$\begin{aligned} F_k(x) = & \{x_{k-1} (D \otimes I) - x_k [\text{diag} (De) \otimes I]\} L_k (x_{k-1}, x_k) \\ & + x_k \{[C + \text{diag} (De)] \oplus T\} + x_{k+1} (I \otimes T^0 \alpha). \end{aligned} \quad (35)$$

Note that $u_0 = g_0 \exp \{(C + D) t\}$ may be regarded as a given vector. Thus $F(x)$ is in $\mathbf{C}^2 (\mathbf{R}_+^\infty)$, and the system of differential vector equations (31) to (33) is rewritten as

$$\frac{d}{dt} x = F(x) \quad (36)$$

with the initial condition

$$x(0) = \tilde{\mathbf{g}} = (g_1, g_2, g_3, \dots). \quad (37)$$

In what follows we show that the expected fraction vector $F(x)$ is Lipschitz.

Based on the definition of the Gateaux derivative, it follows from (34) and (35) that

$$\frac{\partial F(x)}{\partial x} = \begin{pmatrix} \frac{\partial F_1(x)}{\partial x_1} & \frac{\partial F_2(x)}{\partial x_1} & \frac{\partial F_3(x)}{\partial x_1} & & \\ \frac{\partial F_1(x)}{\partial x_2} & \frac{\partial F_2(x)}{\partial x_2} & \frac{\partial F_3(x)}{\partial x_2} & \frac{\partial F_4(x)}{\partial x_2} & \\ & \frac{\partial F_2(x)}{\partial x_3} & \frac{\partial F_3(x)}{\partial x_3} & \frac{\partial F_4(x)}{\partial x_3} & \\ & & \ddots & \ddots & \ddots \end{pmatrix}.$$

We write

$$\mathcal{D}F(x) = \begin{pmatrix} A_1(x) & B_1(x) & & & \\ C_2(x) & A_2(x) & B_2(x) & & \\ & C_3(x) & A_3(x) & B_3(x) & \\ & & \ddots & \ddots & \ddots \end{pmatrix} = \frac{\partial F(x)}{\partial x}. \quad (38)$$

where $A_k(x)$, $B_k(x)$ and $C_j(x)$ are the matrices of size $m_A m_B$ for $k \geq 1$ and $j \geq 2$.

To compute the matrix $\mathcal{D}F(x)$, we need to use two basic properties of the Gateaux derivative as follows:

Property one

$$\frac{\partial x_k}{\partial x_k} = I, \quad \frac{\partial x_k S}{\partial x_k} = S,$$

where S is a matrix of size $m_A m_B$.

Note that

$$L_1(u_0 \otimes \alpha, x_1) = \frac{(u_0 e)^d - (x_1 e)^d}{u_0 e - x_1 e} = \frac{1 - (x_1 e)^d}{1 - x_1 e}$$

and for $k \geq 2$

$$L_k(x_{k-1}, x_k) = \frac{(x_{k-1}e)^d - (x_k e)^d}{x_{k-1}e - x_k e}.$$

Let $y_1 = x_1 e$. Then

$$\begin{aligned} \frac{\partial L_1(u_0 \otimes \alpha, x_1)}{\partial x_1} &= \frac{\partial y_1}{\partial x_1} \frac{\partial L_1(u_0 \otimes \alpha, x_1)}{\partial y_1} \\ &= e \frac{\left[(u_0 e)^d - (x_1 e)^d \right] - d(x_1 e)^{d-1}(u_0 e - x_1 e)}{(u_0 e - x_1 e)^2}. \end{aligned}$$

Similarly, for $k \geq 2$ we can obtain

$$\frac{\partial L_k(x_{k-1}, x_k)}{\partial x_{k-1}} = e \frac{d(x_{k-1}e)^{d-1}(x_{k-1}e - x_k e) - \left[(x_{k-1}e)^d - (x_k e)^d \right]}{(x_{k-1}e - x_k e)^2}$$

and

$$\frac{\partial L_k(x_{k-1}, x_k)}{\partial x_k} = e \frac{\left[(x_{k-1}e)^d - (x_k e)^d \right] - d(x_k e)^{d-1}(x_{k-1}e - x_k e)}{(x_{k-1}e - x_k e)^2}.$$

It is easy to check that

$$\begin{aligned} A_1(x) &= [C + \text{diag}(De)] \oplus T + [\text{diag}(De) \otimes I] \frac{(u_0 e)^d - (x_1 e)^d}{u_0 e - x_1 e} \\ &\quad + e x_1 [\text{diag}(De) \otimes I] \frac{\left[(u_0 e)^d - (x_1 e)^d \right] - d(x_1 e)^{d-1}(u_0 e - x_1 e)}{(u_0 e - x_1 e)^2}, \end{aligned} \quad (39)$$

$$\begin{aligned} B_1(x) &= (D \otimes I) \frac{(x_1 e)^d - (x_2 e)^d}{x_1 e - x_2 e} + e \{x_1 (D \otimes I) - x_2 [\text{diag}(De) \otimes I]\} \\ &\quad \times \frac{d(x_1 e)^{d-1}(x_1 e - x_2 e) - \left[(x_1 e)^d - (x_2 e)^d \right]}{(x_1 e - x_2 e)^2}; \end{aligned} \quad (40)$$

and for $k \geq 2$

$$C_k(x) = I \otimes T^0 \alpha, \quad (41)$$

$$\begin{aligned} B_k(x) &= (D \otimes I) \frac{(x_k e)^d - (x_{k+1} e)^d}{x_k e - x_{k+1} e} + e \{x_k (D \otimes I) - x_{k+1} [\text{diag}(De) \otimes I]\} \\ &\quad \times \frac{d(x_k e)^{d-1}(x_k e - x_{k+1} e) - \left[(x_k e)^d - (x_{k+1} e)^d \right]}{(x_k e - x_{k+1} e)^2}, \end{aligned} \quad (42)$$

$$\begin{aligned} A_k(x) &= [C + \text{diag}(De)] \oplus T + [\text{diag}(De) \otimes I] \frac{(x_{k-1}e)^d - (x_k e)^d}{x_{k-1}e - x_k e} \\ &\quad + e \{x_{k-1} (D \otimes I) - x_k [\text{diag}(De) \otimes I]\} \\ &\quad \times \frac{\left[(x_{k-1}e)^d - (x_k e)^d \right] - d(x_k e)^{d-1}(x_{k-1}e - x_k e)}{(x_{k-1}e - x_k e)^2}. \end{aligned} \quad (43)$$

Note that $\|\mathbf{A}\| = \max_i \left\{ \sum_j |a_{i,j}| \right\}$, it follows from (38) that

$$\|\mathcal{D}F(x)\| = \max \left\{ \|A_1(x)\| + \|B_2(x)\|, \sup_{k \geq 2} \{ \|A_k(x)\| + \|B_k(x)\| + \|C_k(x)\| \} \right\}. \quad (44)$$

Since $u_0e \leq 1$ and $x_1e \leq 1$, we obtain

$$\begin{aligned} \frac{(u_0e)^d - (x_1e)^d}{u_0e - x_1e} &= \sum_{j=0}^{d-1} (u_0e)^j (x_1e)^{d-1-j} \leq d, \\ \frac{\left[(u_0e)^d - (x_1e)^d \right] - d(x_1e)^{d-1}(u_0e - x_1e)}{(u_0e - x_1e)^2} &= \sum_{k=0}^{d-2} \sum_{j=0}^k (u_0e)^j (x_1e)^{k-j} \leq \frac{(d-1)(d-2)}{2}; \\ \frac{(x_{k-1}e)^d - (x_ke)^d}{x_{k-1}e - x_ke} &\leq d, \\ \frac{\left[(x_{k-1}e)^d - (x_ke)^d \right] - d(x_ke)^{d-1}(x_{k-1}e - x_ke)}{(x_{k-1}e - x_ke)^2} &\leq \frac{(d-1)(d-2)}{2}. \end{aligned}$$

Thus it follows from (39) and (40) that

$$\begin{aligned} \|A_1(x)\| &\leq \|C + \text{diag}(De)\| + \frac{2d + (d-1)(d-2)}{2} \|D\| + \|T\|, \\ \|B_1(x)\| &\leq [d + (d-1)(d-2)] \|D\|, \\ \|A_1(x)\| + \|B_1(x)\| &\leq \|C + \text{diag}(De)\| + \left[2d + \frac{3(d-1)(d-2)}{2} \right] \|D\| + \|T\|. \end{aligned}$$

It follows from (41) to (43) that for $k \geq 2$

$$\begin{aligned} \|A_k(x)\| &\leq \|C + \text{diag}(De)\| + [d + (d-1)(d-2)] \|D\| + \|T\|, \\ \|B_k(x)\| &\leq [d + (d-1)(d-2)] \|D\|, \\ \|C_k(x)\| &= \|T^0\alpha\|, \end{aligned}$$

hence we have

$$\begin{aligned} &\|A_k(x)\| + \|B_k(x)\| + \|C_k(x)\| \\ &\leq \|C + \text{diag}(De)\| + 2[d + (d-1)(d-2)] \|D\| + \|T\| + \|T^0\alpha\|. \end{aligned}$$

Let

$$M = \max \left\{ \|C + \text{diag}(De)\| + 2[d + (d-1)(d-2)] \|D\| + \|T\| + \|T^0\alpha\| \right\}.$$

Then

$$\|A_1(x)\| + \|B_1(x)\| \leq M$$

and for $k \geq 2$

$$\|A_k(x)\| + \|B_k(x)\| + \|C_k(x)\| \leq M.$$

Hence, it follows from Equation (44) that

$$\|\mathcal{D}F(x)\| \leq M.$$

Note that $x = \mathbf{u}$, this gives that for $\mathbf{u} \in \tilde{\Omega}$

$$\|\mathcal{D}F(\mathbf{u})\| \leq M. \quad (45)$$

For $\mathbf{u}, \mathbf{v} \in \tilde{\Omega}$,

$$\begin{aligned} \|F(\mathbf{u}) - F(\mathbf{v})\| &\leq \sup_{0 \leq t \leq 1} \|\mathcal{D}F(\mathbf{u} + t(\mathbf{v} - \mathbf{u}))\| \|\mathbf{u} - \mathbf{v}\| \\ &\leq M \|\mathbf{u} - \mathbf{v}\|. \end{aligned} \quad (46)$$

This indicates that the function $F(\mathbf{u})$ is Lipschitz for $\mathbf{u} \in \tilde{\Omega}$.

Note that $x = \mathbf{u}$, it follows from Equations (31) and (33) that for $\mathbf{u} \in \tilde{\Omega}$

$$\mathbf{u}(t) = \mathbf{u}(0) + \int_0^t F(\mathbf{u}(\xi)) d\xi,$$

this gives

$$\mathbf{u}(t) = \tilde{\mathbf{g}} + \int_0^t F(\mathbf{u}(\xi)) d\xi. \quad (47)$$

Using the Picard approximation as well as the Lipschitz condition, it is easy to prove that there exists the unique solution to the integral equation (47) according to the basic results of the Banach space. Therefore, there exists the unique solution to the system of differential vector equations (31) to (33) (that is, (22) to (26)).

6 A Matrix-Analytic Solution

In this section, we first discuss the stability of this supermarket model in terms of a coupling method. Then we provide a generalized matrix-analytic method for computing the fixed point whose doubly exponential solution and phase-structured tail are obtained. Finally, we discuss some useful limits of the fraction vector $\mathbf{u}^{(N)}(t)$ as $N \rightarrow \infty$ and $t \rightarrow +\infty$.

6.1 Stability of this supermarket model

In this subsection, we provide a coupling method to study the stability of this supermarket model of N identical servers with MAP inputs and PH service times, and give a sufficient condition under which this supermarket model is stable.

Let Q and R denote two supermarket models with MAP inputs and PH service times, both of which have the same parameters $N, d, m_A, C, D, m_B, \alpha, T$, and the same initial state at $t = 0$. Let $d(Q)$ and $d(R)$ be two choice numbers in the two supermarket models Q and R , respectively. We assume $d(Q) = 1$ and $d(R) \geq 2$. Thus, the only difference between the two supermarket models Q and R is the two different choice numbers: $d(Q) = 1$ and $d(R) \geq 2$.

For the two supermarket models Q and R , we define two infinite-dimensional Markov processes $\{U_N^{(Q)}(t) : t \geq 0\}$ and $\{U_N^{(R)}(t) : t \geq 0\}$, respectively. The following theorem sets up a coupling between the two processes $\{U_N^{(Q)}(t) : t \geq 0\}$ and $\{U_N^{(R)}(t) : t \geq 0\}$.

Theorem 3 *For the two supermarket models Q and R , there is a coupling between the two processes $\{U_N^{(Q)}(t) : t \geq 0\}$ and $\{U_N^{(R)}(t) : t \geq 0\}$ such that the total number of customers in the supermarket model R is no greater than the total number of customers in the supermarket model Q at time $t \geq 0$.*

Proof: See Appendix C. ■

Remark 3 *Note that the N queueing processes in this supermarket model is symmetric, it is easy to see from Theorem 3 that the queue length of each server in the supermarket model R is no greater than that in the supermarket model Q at time $t \geq 0$.*

Since this supermarket model with MAP inputs and PH service times is more general, it is necessary to extend the coupling method given in Turner [30] and Martin and Suhov [22] through a detailed probability analysis given in Appendix C. We show that such a coupling method can be applied to discussing stability of more general supermarket models.

Note that the stationary arrival rate of the MAP of irreducible matrix descriptor (C, D) is given by $\lambda = \omega D e$, and the mean of the PH service time is given by $1/\mu = -\alpha T^{-1} e$. The following theorem provides a sufficient condition under which this supermarket model is stable.

Theorem 4 *This supermarket model of N identical servers with MAP inputs and PH service times is stable if $\rho = \lambda/\mu < 1$.*

Proof: From the two different choice numbers: $d(Q) = 1$ and $d(R) \geq 2$, we set up two different supermarket models Q and R , respectively. Note that the supermarket model Q is the set of N parallel and independent MAP/PH/1 queues. Obviously, the MAP/PH/1 queue is described as a QBD process whose infinitesimal generator is given by

$$\mathbf{Q} = \begin{pmatrix} C & D \otimes \alpha & & & \\ I \otimes T^0 & C \oplus T & D \otimes I & & \\ & I \otimes (T^0 \alpha) & C \oplus T & D \otimes I & \\ & & \ddots & \ddots & \ddots \end{pmatrix}.$$

Note that

$$A = A_{-1} + A_0 + A_1 = (C + D) \oplus (T + T^0 \alpha),$$

where

$$A_{-1} = I \otimes (T^0 \alpha), \quad A_0 = C \oplus T, \quad A_1 = D \otimes I,$$

thus it is easy to check that $\omega \otimes \theta$ is the stationary probability vector of the Markov chain A , where θ is the stationary probability vector of the Markov chain $T + T^0 \alpha$. Using Chapter 3 of Li [11], it is clear that the QBD process \mathbf{Q} is stable if $(\omega \otimes \theta) A_{-1} e > (\omega \otimes \theta) A_2 e$, that is, $\rho = \lambda/\mu < 1$. Hence, the supermarket model Q is stable if $\rho < 1$. It is seen from Theorem 3 and Remark 3 that the queue length of each server in the supermarket model R is no greater than that in the supermarket model Q at time $t \geq 0$, this shows that the supermarket model R is stable if the supermarket model Q is stable. Thus the supermarket model R is stable if $\rho = \lambda/\mu < 1$. This completes the proof. ■

6.2 Computation of the fixed point

A row vector $\pi = (\pi_0, \pi_1, \pi_2, \dots)$ is called a fixed point of the infinite-dimensional system of differential vector equations (22) to (26) satisfied by the limiting fraction vector $\mathbf{u}(t)$ if $\pi = \lim_{t \rightarrow +\infty} \mathbf{u}(t)$, or $\pi_k = \lim_{t \rightarrow +\infty} u_k(t)$ for $k \geq 0$.

It is well-known that if π is a fixed point of the vector $\mathbf{u}(t)$, then

$$\lim_{t \rightarrow +\infty} \left[\frac{d}{dt} \mathbf{u}(t) \right] = 0.$$

Let

$$L_1(\pi_0 \otimes \alpha, \pi_1) = \sum_{m=1}^d C_d^m \left[\sum_{l=1}^{m_A} \sum_{j=1}^{m_B} (\pi_{0;l} \alpha_j - \pi_{1;l,j}) \right]^{m-1} \left[\sum_{l=1}^{m_A} \sum_{j=1}^{m_B} \pi_{1;l,j} \right]^{d-m}$$

for $k \geq 2$

$$L_k(\pi_{k-1}, \pi_k) = \sum_{m=1}^d C_d^m \left[\sum_{l=1}^{m_A} \sum_{j=1}^{m_B} (\pi_{k-1;l,j} - \pi_{k;l,j}) \right]^{m-1} \left[\sum_{l=1}^{m_A} \sum_{j=1}^{m_B} \pi_{k;l,j} \right]^{d-m}.$$

Then

$$L_1(\pi_0 \otimes \alpha, \pi_1) = \frac{1 - (\pi_1 e)^d}{1 - \pi_1 e}$$

and for $k \geq 2$

$$L_k(\pi_{k-1}, \pi_k) = \frac{(\pi_{k-1} e)^d - (\pi_k e)^d}{\pi_{k-1} e - \pi_k e}.$$

To determine the fixed point $\pi = (\pi_0, \pi_1, \pi_2, \dots)$, as $t \rightarrow +\infty$ taking limits on both sides of Equations (22) to (26) we obtain the system of nonlinear vector equations as follows:

$$\pi_0(C + D) = 0, \quad \pi_0 e = 1, \quad (48)$$

$$\begin{aligned} & \{(\pi_0 \otimes \alpha)(D \otimes I) - \pi_1 [\text{diag}(De) \otimes I]\} L_1(\pi_0 \otimes \alpha, \pi_1) \\ & + \pi_1 \{[C + \text{diag}(De)] \oplus T\} + \pi_2 (I \otimes T^0 \alpha) = 0, \end{aligned} \quad (49)$$

for $k \geq 2$

$$\begin{aligned} & \{\pi_{k-1}(D \otimes I) - \pi_k [\text{diag}(De) \otimes I]\} L_k(\pi_{k-1}, \pi_k) \\ & + \pi_k \{[C + \text{diag}(De)] \oplus T\} + \pi_{k+1} (I \otimes T^0 \alpha) = 0. \end{aligned} \quad (50)$$

Since ω is the stationary probability vector of the Markov chain $C + D$, then it follows from (48) that

$$\pi_0 = \omega. \quad (51)$$

For the fixed point $\pi = (\pi_0, \pi_1, \pi_2, \dots)$, $(\pi_0 e, \pi_1 e, \pi_2 e, \dots)$ is the tail vector of the stationary queue length distribution. The following theorem shows that the tail vector $(\pi_0 e, \pi_1 e, \pi_2 e, \dots)$ of the stationary queue length distribution is doubly exponential.

Theorem 5 *If $\rho = \lambda/\mu < 1$, then the tail vector $(\pi_0 e, \pi_1 e, \pi_2 e, \dots)$ of the stationary queue length distribution is doubly exponential, that is, for $k \geq 0$*

$$\pi_k e = \rho^{\frac{d^k - 1}{d - 1}}. \quad (52)$$

Proof: Multiplying both sides of the equation (50) by the vector e , and noting that $[C + \text{diag}(De)]e = 0$ and $Te = -T^0$, we obtain that

$$[(\pi_0 \otimes \alpha)(De \otimes e) - \pi_1(De \otimes e)]L_k(\pi_0 \otimes, \pi_1) - \mu[\pi_1(e \otimes T^0) - \pi_2(e \otimes T^0)] = 0 \quad (53)$$

for $k \geq 2$,

$$[\pi_{k-1}(De \otimes e) - \pi_k(De \otimes e)]L_k(\pi_{k-1}, \pi_k) - \mu[\pi_k(e \otimes T^0) - \pi_{k+1}(e \otimes T^0)] = 0. \quad (54)$$

Let $\pi_k = \eta_k(\omega \otimes \theta)$ for $k \geq 1$, and $\zeta_1 = L_1(\pi_0 \otimes \alpha, \pi_1)$ and $\zeta_k = L_k(\pi_{k-1}, \pi_k)$ for $k \geq 2$. Note that $\lambda = \omega De$, $\mu = \theta T^0$ and $\rho = \lambda/\mu$, it follows from (54) that

$$\rho(1 - \eta_1^d) - (\eta_1 - \eta_2) = 0$$

and

$$\rho(\eta_{k-1}^d - \eta_k^d) - (\eta_k - \eta_{k+1}) = 0.$$

This gives

$$\pi_k e = \eta_k = \rho^{\frac{d^k - 1}{d - 1}}.$$

This completes the proof. ■

Note that

$$\zeta_k = \frac{\rho^{\frac{d^k - d}{d - 1}} - \rho^{\frac{d^{k+1} - d}{d - 1}}}{\rho^{\frac{d^{k-1} - 1}{d - 1}} - \rho^{\frac{d^k - 1}{d - 1}}}, \quad k \geq 1,$$

we obtain

$$B_k = [C + (1 - \zeta_k) \text{diag}(De)] \oplus T$$

and

$$Q = \begin{pmatrix} B_1 & \zeta_2(D \otimes I) & & & \\ I \otimes (T^0 \alpha) & B_2 & \zeta_3(D \otimes I) & & \\ & I \otimes (T^0 \alpha) & B_3 & \zeta_4(D \otimes I) & \\ & & \ddots & \ddots & \ddots \end{pmatrix}.$$

Then the level-dependent QBD process is irreducible and transient, since

$$\zeta_1 > \zeta_2 > \zeta_3 > \cdots > 0,$$

$$[B_1 + \zeta_2(D \otimes I)]e = -(\zeta_1 - \zeta_2)[(De) \otimes e] - e \otimes T^0 \preceq 0$$

and

$$[I \otimes (T^0 \alpha) + B_k + \zeta_k(D \otimes I)]e = -(\zeta_k - \zeta_{k+1})[(De) \otimes e] \preceq 0.$$

In what follows we provide the UL-type of RG -factorization of the QBD process Q according to Chapter 1 in Li [11] or Li and Cao [14]. Applying the UL-type of RG -Factorization, we can give the maximal non-positive inverse of matrix Q , which leads to the matrix-product solution of the fixed point $(\pi_0, \pi_1, \pi_2, \dots)$ by means of the R - and U -measures.

Let the matrix sequence $\{R_k, k \geq 1\}$ be the minimal nonnegative solution to the nonlinear matrix equations

$$\xi_{k+1} (D \otimes I) + R_k B_{k+1} + R_k R_{k+1} [I \otimes (T^0 \alpha)] = 0,$$

and the matrix sequence $\{G_k, k \geq 2\}$ be the minimal nonnegative solution to the nonlinear matrix equations

$$I \otimes (T^0 \alpha) + B_k G_k + \xi_{k+1} (D \otimes I) G_{k+1} G_k = 0.$$

Let the matrix sequence $\{U_k, k \geq 0\}$ be

$$\begin{aligned} U_k &= B_{k+1} + [\zeta_{k+2} (D \otimes I)] [-U_{k+1}]^{-1} [I \otimes (T^0 \alpha)] \\ &= B_{k+1} + R_{k+1} [I \otimes (T^0 \alpha)] \\ &= B_{k+1} + [\zeta_{k+2} (D \otimes I)] G_{k+1}. \end{aligned}$$

Hence we obtain

$$R_0 = \zeta_1 (D \otimes I) (-U_1)^{-1}$$

and

$$G_1 = (-U_0)^{-1} [I \otimes (T^0 \alpha)].$$

Based on the R -measure $\{R_k, k \geq 0\}$, G -measure $\{G_k, k \geq 1\}$ and U -measure $\{U_k, k \geq 0\}$, we can get the UL-type of RG -factorization of the matrix Q as follows

$$Q = (I - R_U) U_D (I - G_L),$$

where

$$R_U = \begin{pmatrix} 0 & R_0 & & & \\ & 0 & R_1 & & \\ & & 0 & R_2 & \\ & & & \ddots & \ddots \end{pmatrix},$$

$$U_D = \text{diag}(U_0, U_1, U_2, \dots)$$

and

$$G_L = \begin{pmatrix} I & & & & \\ G_1 & I & & & \\ & G_2 & I & & \\ & & \ddots & \ddots & \ddots \end{pmatrix}.$$

Using the RG -factorization, we obtain the maximal non-positive inverse of the matrix Q as follows

$$Q^{-1} = (I - G_L)^{-1} U_D^{-1} (I - R_U)^{-1}, \quad (55)$$

where

$$(I - R_U)^{-1} = \begin{pmatrix} I & X_1^{(0)} & X_2^{(0)} & X_3^{(0)} & \cdots \\ & I & X_1^{(1)} & X_2^{(1)} & \cdots \\ & & I & X_1^{(2)} & \cdots \\ & & & I & \cdots \\ & & & & \ddots \end{pmatrix},$$

$$X_k^{(l)} = R_l R_{l+1} R_{l+2} \cdots R_{l+k-1}, \quad k \geq 1, l \geq 0;$$

$$U_D^{-1} = \text{diag}(U_0^{-1}, U_1^{-1}, U_2^{-1}, \dots);$$

$$(I - G_L)^{-1} = \begin{pmatrix} I & & & & \\ Y_1^{(1)} & I & & & \\ Y_2^{(2)} & Y_1^{(2)} & I & & \\ Y_3^{(3)} & Y_2^{(3)} & Y_1^{(3)} & I & \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

$$Y_k^{(l)} = G_l G_{l-1} G_{l-2} \cdots G_{l-k+1}, \quad l \geq k \geq 1.$$

The following theorem illustrates that the fixed point $(\pi_0, \pi_1, \pi_2, \dots)$ is matrix-product.

Theorem 6 *If $\rho < 1$, then the fixed point $\pi = (\pi_0, \pi_1, \pi_2, \dots)$ is given by*

$$\pi_0 = \omega,$$

$$\pi_1 = \zeta_1(\omega \otimes \alpha)(D \otimes I)(-U_0)^{-1} \quad (56)$$

and for $k \geq 2$

$$\pi_k = \zeta_1(\omega \otimes \alpha)(D \otimes I)(-U_0)^{-1} R_0 R_1 \cdots R_{k-2}. \quad (57)$$

Proof: It follows from (54) that

$$\begin{aligned}
& (\pi_1, \pi_2, \pi_3, \dots) \begin{pmatrix} B_1 & \zeta_2(D \otimes I) & & \\ I \otimes (T^0 \alpha) & B_2 & \zeta_3(D \otimes I) & \\ & I \otimes (T^0 \alpha) & B_3 & \zeta_4(D \otimes I) \\ & & \ddots & \ddots & \ddots \end{pmatrix} \\
& = -(\zeta_1(\omega \otimes \alpha)(D \otimes I), 0, 0, \dots).
\end{aligned}$$

This gives

$$(\pi_1, \pi_2, \pi_3, \dots) = -(\zeta_1(\omega \otimes \alpha)(D \otimes I), 0, 0, \dots)(I - G_L)^{-1} U_D^{-1} (I - R_U)^{-1}.$$

Thus we obtain

$$\pi_1 = \zeta_1(\omega \otimes \alpha)(D \otimes I)(-U_0)^{-1}$$

and for $k \geq 2$

$$\pi_k = \zeta_1(\omega \otimes \alpha)(D \otimes I)(-U_0)^{-1} R_0 R_1 \cdots R_{k-2}.$$

This completes the proof. ■

In what follows we consider the block-structured supermarket model with Poisson inputs and PH service times. In this case, we can give an interesting explicit expression of the fixed point.

Note that $C = -\lambda$, $D = \lambda$, it is clear that $\omega = 1$ and $\pi_0 = 1$. It follows from Equations (49) and (50) that

$$\lambda(\theta - \pi_1) \frac{1 - (\pi_1 e)^d}{1 - (\pi_1 e)} + \pi_1 T + \pi_2 T^0 \alpha = 0$$

and for $k \geq 2$

$$\lambda(\pi_{k-1} - \pi_k) \frac{(\pi_{k-1} e)^d - (\pi_k e)^d}{(\pi_{k-1} e) - (\pi_k e)} + \pi_k T + \pi_{k+1} T^0 \alpha = 0.$$

Thus we obtain

$$(\pi_1, \pi_2, \pi_3, \dots) \Theta = \lambda \left((\theta - \pi_1) \frac{1 - (\pi_1 e)^d}{1 - (\pi_1 e)}, (\pi_1 - \pi_2) \frac{(\pi_1 e)^d - (\pi_2 e)^d}{(\pi_1 e) - (\pi_2 e)}, \dots \right), \quad (58)$$

where

$$\Theta = \begin{pmatrix} -T & & & \\ -T^0 \alpha & -T & & \\ & -T^0 \alpha & -T & \\ & & \ddots & \ddots \end{pmatrix}.$$

Since

$$\Theta^{-1} = \begin{pmatrix} (-T)^{-1} & & & & \\ (e\alpha)(-T)^{-1} & (-T)^{-1} & & & \\ (e\alpha)(-T)^{-1} & (e\alpha)(-T)^{-1} & (-T)^{-1} & & \\ (e\alpha)(-T)^{-1} & (e\alpha)(-T)^{-1} & (e\alpha)(-T)^{-1} & (-T)^{-1} & \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

It follows from (58) that

$$\pi_1 \left[I + \lambda \frac{1 - (\pi_1 e)^d}{1 - (\pi_1 e)} (-T)^{-1} \right] = \lambda \frac{1 - (\pi_1 e)^d}{1 - (\pi_1 e)} \theta (-T)^{-1} + \lambda \alpha (-T)^{-1} (\pi_1 e)^d \quad (59)$$

and for $k \geq 2$

$$\pi_k \left[I + \lambda \frac{(\pi_{k-1} e)^d - (\pi_k e)^d}{(\pi_{k-1} e) - (\pi_k e)} (-T)^{-1} \right] = \lambda \frac{(\pi_{k-1} e)^d - (\pi_k e)^d}{(\pi_{k-1} e) - (\pi_k e)} \theta (-T)^{-1} + \lambda \alpha (-T)^{-1} (\pi_k e)^d. \quad (60)$$

Note that the matrices $I + \lambda \frac{1 - (\pi_1 e)^d}{1 - (\pi_1 e)} (-T)^{-1}$ and $I + \lambda \frac{(\pi_{k-1} e)^d - (\pi_k e)^d}{(\pi_{k-1} e) - (\pi_k e)} (-T)^{-1}$ for $k \geq 2$ are all invertible, it follows from (59) and (60) that

$$\pi_1 = \left[\lambda \frac{1 - (\pi_1 e)^d}{1 - (\pi_1 e)} \omega (-T)^{-1} + \lambda \alpha (-T)^{-1} (\pi_1 e)^d \right] \left[I + \lambda \frac{1 - (\pi_1 e)^d}{1 - (\pi_1 e)} (-T)^{-1} \right]^{-1}.$$

and for $k \geq 2$

$$\begin{aligned} \pi_k &= \left[\lambda \frac{(\pi_{k-1} e)^d - (\pi_k e)^d}{(\pi_{k-1} e) - (\pi_k e)} \omega (-T)^{-1} + \lambda \alpha (-T)^{-1} (\pi_k e)^d \right] \\ &\quad \times \left[I + \lambda \frac{(\pi_{k-1} e)^d - (\pi_k e)^d}{(\pi_{k-1} e) - (\pi_k e)} (-T)^{-1} \right]^{-1}. \end{aligned}$$

Thus we obtain

$$\pi_1 = \left[\lambda \zeta_1 \omega (-T)^{-1} + \lambda \alpha (-T)^{-1} \rho^d \right] \left[I + \lambda \zeta_1 (-T)^{-1} \right]^{-1} \quad (61)$$

and for $k \geq 2$

$$\pi_k = \left[\lambda \zeta_k \omega (-T)^{-1} + \lambda \alpha (-T)^{-1} \rho^{\frac{d^{k+1}-d}{d-1}} \right] \left[I + \lambda \zeta_k (-T)^{-1} \right]^{-1}. \quad (62)$$

Remark 4 For this block-structured supermarket model, the fixed point is matrix-product and depends on the R -measure $\{R_k, k \geq 0\}$, see (56) and (57). However, when the input is a Poisson process, we can give the explicit expression of the fixed point by (61) and (62). This explains the reason why the MAP input makes the study of block-structured supermarket models more difficult and challenging.

6.3 The double limits

In this subsection, we discuss some useful limits of the fraction vector $\mathbf{u}^{(N)}(t)$ as $N \rightarrow \infty$ and $t \rightarrow +\infty$. Note that the limits are necessary for using the stationary probabilities of the limiting process to give an effective approximate performance of this supermarket model.

The following theorem gives the limit of the vector $\mathbf{u}(t, \mathbf{g})$ as $t \rightarrow +\infty$, that is,

$$\lim_{t \rightarrow +\infty} \mathbf{u}(t, \mathbf{g}) = \lim_{t \rightarrow +\infty} \lim_{N \rightarrow \infty} \mathbf{u}^{(N)}(t, \mathbf{g}).$$

Theorem 7 *If $\rho < 1$, then for any $\mathbf{g} \in \Omega$*

$$\lim_{t \rightarrow +\infty} \mathbf{u}(t, \mathbf{g}) = \pi.$$

Furthermore, there exists a unique probability measure φ on Ω , which is invariant under the map $\mathbf{g} \mapsto \mathbf{u}(t, \mathbf{g})$, that is, for any continuous function $f : \Omega \rightarrow \mathbf{R}$ and $t > 0$

$$\int_{\Omega} f(\mathbf{g}) d\varphi(\mathbf{g}) = \int_{\Omega} f(\mathbf{u}(t, \mathbf{g})) d\varphi(\mathbf{g}).$$

Also, $\varphi = \delta_{\pi}$ is the probability measure concentrated at the fixed point π .

Proof: It is seen from Theorem 6 that the condition $\rho < 1$ guarantees the existence of solution in Ω to the system of nonlinear equations (48) to (50). This indicates that if $\rho < 1$, then as $t \rightarrow +\infty$, the limit of $\mathbf{u}(t, \mathbf{g})$ exists in Ω . Since $\mathbf{u}(t, \mathbf{g})$ is the unique and global solution to the infinite-dimensional system of differential vector equations (22) to (26) for $t \geq 0$, the vector $\lim_{t \rightarrow +\infty} \mathbf{u}(t, \mathbf{g})$ is also a solution to the system of nonlinear equations (48) to (50). Note that π is the unique solution to the system of nonlinear equations (48) to (50), hence we obtain that $\lim_{t \rightarrow +\infty} \mathbf{u}(t, \mathbf{g}) = \pi$. The second statement in this theorem can be immediately given by the probability measure of the limiting process $\{U(t), t \geq 0\}$ on state space Ω . This completes the proof. ■

The following theorem indicates the weak convergence of the sequence $\{\varphi_N\}$ of stationary probability distributions for the sequence $\{U^{(N)}(t), t \geq 0\}$ of Markov processes to the probability measure concentrated at the fixed point π .

Theorem 8 (1) *If $\rho < 1$, then for a fixed number $N = 1, 2, 3, \dots$, the Markov process $\{U^{(N)}(t), t \geq 0\}$ is positive recurrent, and has a unique invariant distribution φ_N .*

(2) *$\{\varphi_N\}$ weakly converges to δ_{π} , that is, for any continuous function $f : \Omega \rightarrow \mathbf{R}$*

$$\lim_{N \rightarrow \infty} E_{\varphi_N} [f(\mathbf{g})] = f(\pi).$$

Proof: (1) From Theorem 3, this supermarket model of N identical servers is stable if $\rho < 1$, hence this supermarket model has a unique invariant distribution φ_N .

(2) Since $\tilde{\Omega}$ is compact under the metric $\rho(\mathbf{u}, \mathbf{u}')$ given in (67), so is the set $\mathcal{P}(\tilde{\Omega})$ of probability measures. Hence the sequence $\{\varphi_N\}$ of invariant distributions has limiting points. A similar analysis to the proof of Theorem 5 in Martin and Suhov [22] shows that $\{\varphi_N\}$ weakly converges to δ_π and $\lim_{N \rightarrow \infty} E_{\varphi_N}[f(\mathbf{g})] = f(\pi)$. This completes the proof. ■

Based on Theorems 7 and 8, we obtain a useful relation as follows

$$\lim_{t \rightarrow +\infty} \lim_{N \rightarrow \infty} \mathbf{u}^{(N)}(t, \mathbf{g}) = \lim_{N \rightarrow \infty} \lim_{t \rightarrow +\infty} \mathbf{u}^{(N)}(t, \mathbf{g}) = \pi.$$

Therefore, we have

$$\lim_{\substack{N \rightarrow \infty \\ t \rightarrow +\infty}} \mathbf{u}^{(N)}(t, \mathbf{g}) = \pi,$$

which justifies the interchange of the limits of $N \rightarrow \infty$ and $t \rightarrow +\infty$. This is necessary in many practical applications when using the stationary probabilities of the limiting process to give an effective approximate performance of this supermarket model.

7 Performance Computation

In this section, we provide two performance measures of this supermarket model, and use some numerical examples to show how the two performance measures of this supermarket model depend on the non-Poisson MAP inputs and on the non-exponential PH service times.

7.1 Performance measures

For this supermarket model, we provide two simple performance measures as follows:

(1) The mean of the stationary queue length in any server

The mean of the stationary queue length in any server is given by

$$E[Q_d] = \sum_{k=1}^{\infty} \pi_k e = \sum_{k=1}^{\infty} \rho^{\frac{d^k-1}{d-1}}. \quad (63)$$

(2) The expected sojourn time that any arriving customer spends in this system

Note that $u_0^{(N)}(0) \geq 0$ and $u_0^{(N)}(0)e = 1$, it is clear that

$$\lim_{t \rightarrow +\infty} u_0^{(N)}(t) = \lim_{t \rightarrow +\infty} u_0^{(N)}(0) \exp\{(C+D)t\} = \omega.$$

For the PH service times, any arriving customer finds k customer in any server whose probability is given by $(\omega \otimes \alpha - \pi_1) L_d(\omega \otimes \alpha, \pi_1) e$ for $k = 0$ and $(\pi_k - \pi_{k+1}) L_d(\pi_k, \pi_{k+1}) e$ for $k \geq 1$. When $k \geq 1$, the head customer in the server has been served, and so its service time is residual and is denoted as X_R . Let X be of phase type with irreducible representation (α, T) . Then X_R is also of phase type with irreducible representation (θ, T) , where θ is the stationary probability vector of the Markov chain $T + T^0\alpha$. Clearly, we have

$$E[X] = \alpha(-T)^{-1}e, \quad E[X_R] = \theta(-T)^{-1}e.$$

Thus it is easy to see that the expected sojourn time that any arriving customer spends in this system is given by

$$\begin{aligned} E[T_d] &= (\omega \otimes \alpha - \pi_1) L_d(\omega \otimes \alpha, \pi_1) e E[X] \\ &\quad + \sum_{k=1}^{\infty} (\pi_k - \pi_{k+1}) L_d(\pi_k, \pi_{k+1}) e \{E[X_R] + kE[X]\} \\ &= (1 - \rho) E[X] + \sum_{k=1}^{\infty} \left(\rho^{\frac{d^k-1}{d-1}} - \rho^{\frac{d^{k+1}-1}{d-1}} \right) \{E[X_R] + kE[X]\} \\ &= E[X] + \rho E[X_R] + E[X] \sum_{k=2}^{\infty} \rho^{\frac{d^k-1}{d-1}}. \end{aligned} \tag{64}$$

From (63) and (64), we obtain

$$E[T_d] = E[X] E[Q_d] + \rho \{E[X_R] - E[X]\}. \tag{65}$$

Specifically, if $E[X_R] = E[X]$ (for example, the exponential service times), then

$$E[T_d] = E[X] E[Q_d], \tag{66}$$

which is the Little's formula in this supermarket model.

It is seen from (63) that $E[Q_d]$ only depends on the traffic intensity $\rho = \lambda/\mu$, where $\lambda = \omega D e$ and $\mu = -\alpha T^{-1}e$; and from (64) that $E[T_d]$ depends not only on the traffic intensity ρ but also on the mean $E[X_R]$ of the residual PH service time, where $E[X_R] = \theta(-T)^{-1}e$. Based on this, it is clear that performance numerical computation of this supermarket model can be given easily for more general MAP inputs and PH service times, although here our numerical examples are simple.

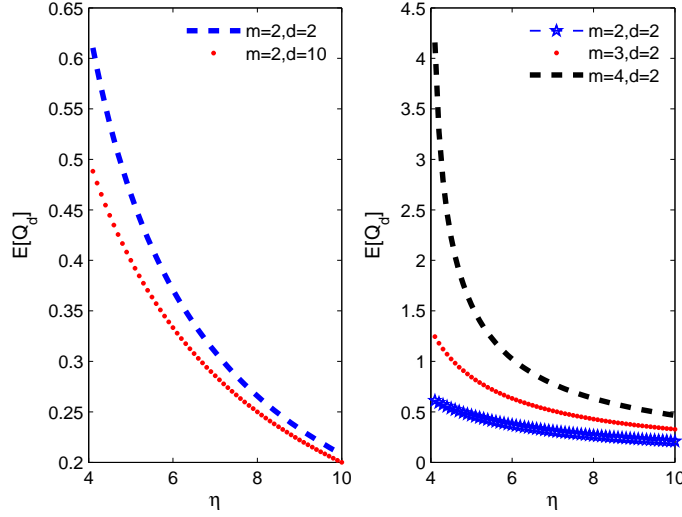


Figure 3: $E[Q_d]$ vs η for $(m, d) = (2, 2), (3, 2), (4, 2)$ and $(2, 10)$

7.2 Numerical examples

In this subsection, we provide some numerical examples which are used to indicate how the performance measures of this supermarket model depend on the non-Poisson MAP inputs and on the non-exponential PH service times.

Example one: The Erlang service times

In this supermarket model, the customers arrive at this system as a Poisson process with arrival rate $N\lambda$, and the service times at each server are an Erlang distribution $E[m, \eta]$. Let $\lambda = 1$. Then $\rho = m/\eta$. When $\rho < 1$, we have $\eta > m$. Figure 3 shows how $E[Q_d]$ depends on the different parameter pairs $(m, d) = (2, 2), (3, 2), (4, 2)$ and $(2, 10)$, respectively. It is seen that $E[Q_d]$ decreases as d increases or as η increases, and it increases as m increases.

Example two: Performance comparisons between the exponential and PH service times

We consider two related supermarket models with Poisson inputs of arrival rate $N\lambda$: one with exponential service times, and another with PH service times. For the two supermarket models, our goal is to observe the influence of different service time distributions on the performance of this supermarket model. To that end, the parameters of this system

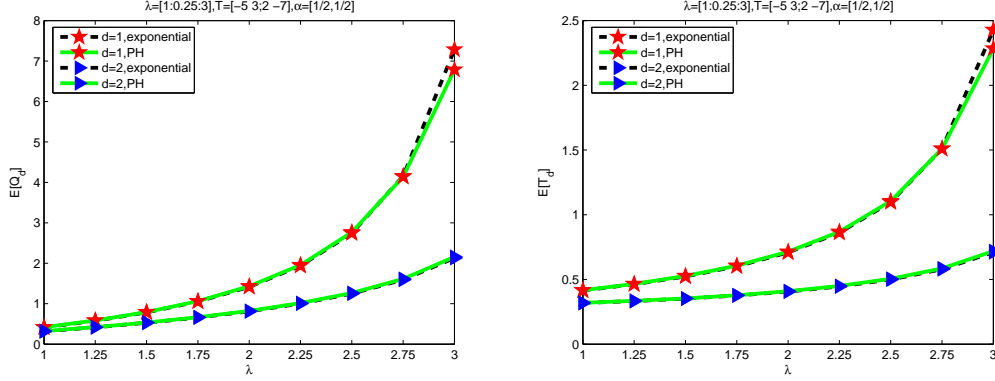


Figure 4: Performance comparison between the exponential and PH service times

are taken as

$$\mu = 3.4118, \quad \alpha = \left(\frac{1}{2}, \frac{1}{2} \right), \quad T = \begin{pmatrix} -5 & 3 \\ 2 & -7 \end{pmatrix}.$$

Under the exponential and PH service times, Figure 4 depicts how $E[Q_d]$ and $E[T_d]$ depend on the arrival rate $\lambda \in [1, 3]$ with $\lambda < \mu$, and on the choice number $d = 1, 2$. It is seen that $E[Q_d]$ and $E[T_d]$ decrease as d increases, while $E[Q_d]$ and $E[T_d]$ increase as λ increases.

Example three: The role of the PH service times

In this supermarket model with $d = 2$, the customers arrive at this system as a Poisson process with arrival rate $N\lambda$, and the service times at each server are a PH distribution with irreducible representation $(\alpha, T(i))$, $\alpha = (1/2, 1/2)$,

$$T(1) = \begin{pmatrix} -5 & 3 \\ 2 & -7 \end{pmatrix}, \quad T(2) = \begin{pmatrix} -4 & 3 \\ 2 & -7 \end{pmatrix}, \quad T(3) = \begin{pmatrix} -4 & 4 \\ 2 & -7 \end{pmatrix}.$$

It is seen that some minor changes are designed in the first rows of the matrices $T(i)$ for $i = 1, 2, 3$. Let $\lambda = 1$. Then

$$\rho(1) = 0.2931, \quad \rho(2) = 0.3636, \quad \rho(3) = 0.4250.$$

This gives

$$\rho(1) < \rho(2) < \rho(3).$$

Figure 5 indicates how $E[T_d]$ depends on the different transition rate matrices $T(i)$ for $i = 1, 2, 3$, and

$$E[T_d(1)] < E[T_d(2)] < E[T_d(3)].$$

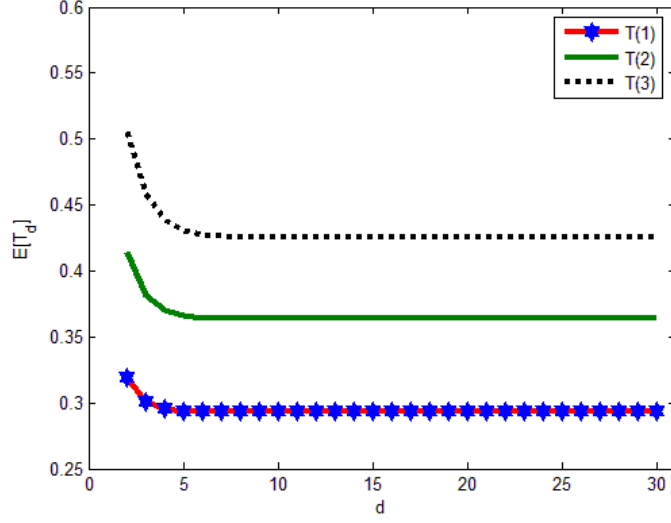


Figure 5: $E[T_d(i)]$ vs the transition rate matrices $T(i)$ for $i = 1, 2, 3$

It is seen that $E[T_d]$ decreases as d increases.

Example four: The role of the MAP inputs

In this supermarket model, the service time distribution is exponential with service rate $\mu = 1$, and the arrival processes are the MAP of irreducible matrix descriptor $(C(N), D(N))$, where

$$C = \begin{pmatrix} -5 - \frac{2}{7}\lambda & 5 \\ 7 & -7 - 2\lambda \end{pmatrix}, \quad D = \begin{pmatrix} \frac{2}{7}\lambda & 0 \\ 0 & 2\lambda \end{pmatrix}.$$

It is easy to check that $\omega = (7/12, 5/12)$, and the stationary arrival rate $\lambda^* = \omega D e = \lambda$. If $\mu = 1$ and $\rho = \lambda^*/\mu = \lambda < 1$, then $\lambda \in (0, 1)$.

Figure 6 shows how $E[Q_d]$ and $E[T_d]$ depend on the parameter λ of the MAP under different choice numbers $d = 1, 2, 5, 10$. It is seen that $E[Q_d]$ and $E[T_d]$ decrease as d increases, while $E[Q_d]$ and $E[T_d]$ increase as λ increases.

8 Concluding Remarks

In this paper, we analyze a more general block-structured supermarket model with non-Poisson MAP inputs and with non-exponential PH service times, and set up an infinite-dimensional system of differential vector equations satisfied by the expected fraction vector

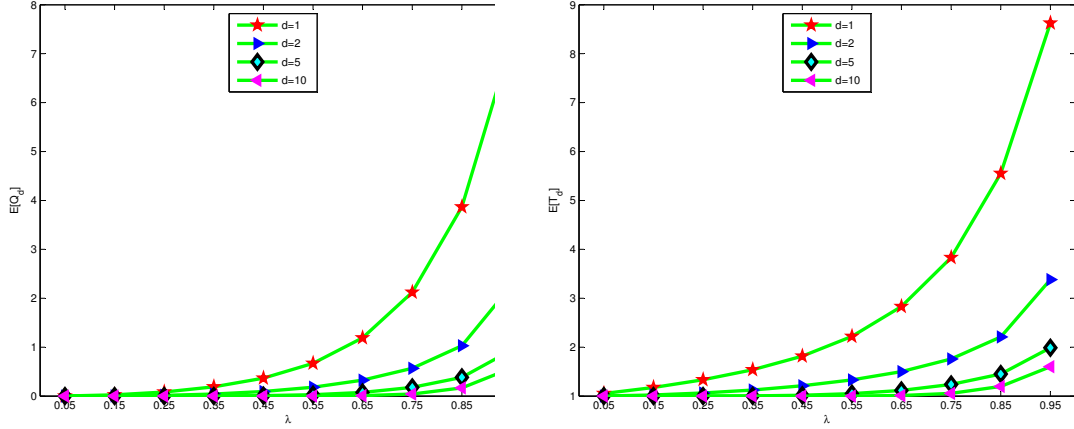


Figure 6: The role of the MAP inputs

through a detailed probability analysis, where an important result: The invariance of environment factors is obtained. We apply the phase-structured operator semigroup to proving the phase-structured mean-field limit, which indicates the asymptotic independence of the block-structured queueing processes in this supermarket model. Furthermore, we provide an effective algorithm for computing the fixed point by means of the matrix-analytic method. Using the fixed point, we provide two performance measures of this supermarket model, and use some numerical examples to illustrate how the two performance measures depend on the non-Poisson MAP inputs and on the non-exponential PH service times. From many practical applications, the block-structured supermarket model is an important queueing model to analyze the relation between the system performance and the job routing rule, and it can also help to design reasonable architecture to improve the performance and to balance the load.

Note that this paper provide a clear picture for how to use the phase-structured mean-field model as well as the matrix-analytic method to analyze performance measures of more general supermarket models. We show that this picture is organized as three key parts: (1) Setting up system of differential equations, (2) necessary proofs of the phase-structured mean-field limit, and (3) performance computation of this supermarket model through the fixed point. Therefore, the results of this paper give new highlight on understanding performance analysis and nonlinear Markov processes for more general supermarket models with non-Poisson inputs and with non-exponential service times. Along such a line, there are a number of interesting directions for potential future research, for example:

- analyzing non-Poisson inputs such as renewal processes;
- studying non-exponential service time distributions, for example, general distributions, matrix-exponential distributions and heavy-tailed distributions; and
- discussing the bulk arrival processes, such as BMAP inputs, and the bulk service processes, where effective algorithms for the fixed point are necessary and interesting.

Up to now, we believe that a larger gap exists when dealing with either renewal inputs or general service times in a supermarket model, because a more challenging infinite-dimensional system of differential equations need be established, a more complicated mean-field limit need be proved, and computation of the fixed point will be more interesting, difficult and challenging.

Acknowledgements

The authors thank the Associate Editor and two reviewers for many valuable comments to sufficiently improve the presentation of this paper. At the same time, the first author acknowledges that this research is partly supported by the National Natural Science Foundation of China (No. 71271187) and the Hebei Natural Science Foundation of China (No. A2012203125).

Three Appendices

Appendix A: Proof of Theorem 1

To prove Equations (9) to (12) in Theorem 1, we need the following computational steps. Note that

$$\begin{aligned}
& \sum_{m=1}^d C_d^m \left\{ \sum_{j=1}^{m_B} \left[u_{k-1;l,j}^{(N)}(t) - u_{k;l,j}^{(N)}(t) \right] \right\}^{m-1} \left\{ \sum_{j=1}^{m_B} \left[u_{k;l,j}^{(N)}(t) \right] \right\}^{d-m} \\
&= C_d^d \left\{ \sum_{j=1}^{m_B} \left[u_{k-1;l,j}^{(N)}(t) - u_{k;l,j}^{(N)}(t) \right] \right\}^{d-1} \\
&+ \sum_{m=1}^{d-1} C_d^m \left\{ \sum_{j=1}^{m_B} \left[u_{k-1;l,j}^{(N)}(t) - u_{k;l,j}^{(N)}(t) \right] \right\}^{m-1} \left\{ \sum_{j=1}^{m_B} \left[u_{k;l,j}^{(N)}(t) \right] \right\}^{d-m}
\end{aligned}$$

and

$$\begin{aligned}
& \sum_{m=1}^{d-1} C_d^m \left\{ \sum_{j=1}^{m_B} \left[u_{k-1;l,j}^{(N)}(t) - u_{k;l,j}^{(N)}(t) \right] \right\}^{m-1} \left\{ \sum_{j=1}^{m_B} \left[u_{k;l,j}^{(N)}(t) \right] \right\}^{d-m} \\
& + \sum_{m=1}^{d-1} C_d^m \left\{ \sum_{j=1}^{m_B} \left[u_{k-1;l,j}^{(N)}(t) - u_{k;l,j}^{(N)}(t) \right] \right\}^{m-1} \\
& \times \sum_{\substack{r_1+r_2+\dots+r_{m_A}=d-m \\ \sum_{i \neq l}^{m_A} r_i \geq 1 \\ 0 \leq r_j \leq d-m, 1 \leq j \leq m_A}} \binom{d-m}{r_1, r_2, \dots, r_{m_A}} \prod_{i=1}^{m_A} \left\{ \sum_{j=1}^{m_B} \left[u_{k;i,j}^{(N)}(t) \right] \right\}^{r_i} \\
& = \sum_{m=1}^{d-1} C_d^m \left\{ \sum_{j=1}^{m_B} \left[u_{k-1;l,j}^{(N)}(t) - u_{k;l,j}^{(N)}(t) \right] \right\}^{m-1} \left\{ \sum_{i=1}^{m_A} \sum_{j=1}^{m_B} \left[u_{k;i,j}^{(N)}(t) \right] \right\}^{d-m},
\end{aligned}$$

since $\left\{ \sum_{j=1}^{m_B} \left[u_{k;l,j}^{(N)}(t) \right] \right\}^{d-m}$ corresponds to the case with $\sum_{i \neq l}^{m_A} r_i = 0$ and $r_l = d-m$, and

$$\begin{aligned}
& \left\{ \sum_{j=1}^{m_B} \left[u_{k;l,j}^{(N)}(t) \right] \right\}^{d-m} + \sum_{\substack{r_1+r_2+\dots+r_{m_A}=d-m \\ \sum_{i \neq l}^{m_A} r_i \geq 1 \\ 0 \leq r_j \leq d-m, 1 \leq j \leq m_A}} \binom{d-m}{r_1, r_2, \dots, r_{m_A}} \prod_{i=1}^{m_A} \left\{ \sum_{j=1}^{m_B} \left[u_{k;i,j}^{(N)}(t) \right] \right\}^{r_i} \\
& = \sum_{\substack{r_1+r_2+\dots+r_{m_A}=d-m \\ 0 \leq r_j \leq d-m, 1 \leq j \leq m_A}} \binom{d-m}{r_1, r_2, \dots, r_{m_A}} \prod_{i=1}^{m_A} \left\{ \sum_{j=1}^{m_B} \left[u_{k;i,j}^{(N)}(t) \right] \right\}^{r_i} \\
& = \left\{ \sum_{i=1}^{m_A} \sum_{j=1}^{m_B} \left[u_{k;i,j}^{(N)}(t) \right] \right\}^{d-m}.
\end{aligned}$$

we obtain

$$\begin{aligned}
& C_d^d \left\{ \sum_{j=1}^{m_B} \left[u_{k-1;l,j}^{(N)}(t) - u_{k;l,j}^{(N)}(t) \right] \right\}^{d-1} \\
& + \sum_{m=1}^{d-1} C_d^m \left\{ \sum_{j=1}^{m_B} \left[u_{k-1;l,j}^{(N)}(t) - u_{k;l,j}^{(N)}(t) \right] \right\}^{m-1} \left\{ \sum_{i=1}^{m_A} \sum_{j=1}^{m_B} u_{k;l,j}^{(N)}(t) \right\}^{d-m} \\
& = \sum_{m=1}^d C_d^m \left\{ \sum_{j=1}^{m_B} \left[u_{k-1;l,j}^{(N)}(t) - u_{k;l,j}^{(N)}(t) \right] \right\}^{m-1} \left\{ \sum_{i=1}^{m_A} \sum_{j=1}^{m_B} \left[u_{k;l,j}^{(N)}(t) \right] \right\}^{d-m} \\
& = C_d^1 \left\{ \sum_{i=1}^{m_A} \sum_{j=1}^{m_B} \left[u_{k;l,j}^{(N)}(t) \right] \right\}^{d-1} + \sum_{m=2}^d C_d^m \left\{ \sum_{j=1}^{m_B} \left[u_{k-1;l,j}^{(N)}(t) - u_{k;l,j}^{(N)}(t) \right] \right\}^{m-1} \\
& \times \left\{ \sum_{i=1}^{m_A} \sum_{j=1}^{m_B} \left[u_{k;l,j}^{(N)}(t) \right] \right\}^{d-m}.
\end{aligned}$$

Using $\frac{m_1}{m} C_m^{m_1} = C_{m-1}^{m_1-1}$, we can obtain

$$\begin{aligned}
& \sum_{m=2}^d C_d^m \left\{ \sum_{j=1}^{m_B} \left[u_{k-1;l,j}^{(N)}(t) - u_{k;l,j}^{(N)}(t) \right] \right\}^{m-1} \left\{ \sum_{i=1}^{m_A} \sum_{j=1}^{m_B} \left[u_{k;l,j}^{(N)}(t) \right] \right\}^{d-m} \\
& + \sum_{m=2}^d C_d^m \sum_{m_1=1}^{m-1} \frac{m_1}{m} C_{m-1}^{m_1-1} \left\{ \sum_{j=1}^{m_B} \left[u_{k-1;l,j}^{(N)}(t) - u_{k;l,j}^{(N)}(t) \right] \right\}^{m_1-1} \\
& \times \sum_{\substack{n_1+n_2+\dots+n_{m_A}=m-m_1 \\ \sum_{i \neq l}^{m_A} n_i \geq 1 \\ 0 \leq n_j \leq m-m_1, 1 \leq j \leq m_A}} \binom{m-m_1}{n_1, n_2, \dots, n_{m_A}} \\
& \times \prod_{i=1}^{m_A} \left\{ \sum_{j=1}^{m_B} \left[u_{k-1;i,j}^{(N)}(t) - u_{k;i,j}^{(N)}(t) \right] \right\}^{n_i} \\
& \times \sum_{\substack{r_1+r_2+\dots+r_{m_A}=d-m \\ 0 \leq r_j \leq d-m, 1 \leq j \leq m_A}} \binom{d-m}{r_1, r_2, \dots, r_{m_A}} \prod_{i=1}^{m_A} \left\{ \sum_{j=1}^{m_B} u_{k;i,j}^{(N)}(t) \right\}^{r_i}
\end{aligned}$$

$$\begin{aligned}
&= \sum_{m=2}^d C_d^m \sum_{m_1=1}^m C_{m-1}^{m_1-1} \left\{ \sum_{j=1}^{m_B} \left[u_{k-1;l,j}^{(N)}(t) - u_{k;l,j}^{(N)}(t) \right] \right\}^{m_1-1} \\
&\times \sum_{\substack{n_1+n_2+\dots+n_{m_A}=m-m_1 \\ 0 \leq n_j \leq m-m_1, 1 \leq j \leq m_A}} \binom{m-m_1}{n_1, n_2, \dots, n_{m_A}} \\
&\times \prod_{i=1}^{m_A} \left\{ \sum_{j=1}^{m_B} \left[u_{k-1;i,j}^{(N)}(t) - u_{k;i,j}^{(N)}(t) \right] \right\}^{n_i} \left\{ \sum_{i=1}^{m_A} \sum_{j=1}^{m_B} \left[u_{k;l,j}^{(N)}(t) \right] \right\}^{d-m} \\
&= \sum_{m=2}^d C_d^m \sum_{m_1=1}^m \frac{m_1}{m} C_{m-1}^{m_1} \left\{ \sum_{j=1}^{m_B} \left[u_{k-1;l,j}^{(N)}(t) - u_{k;l,j}^{(N)}(t) \right] \right\}^{m_1-1} \\
&\times \left\{ \sum_{i \neq l}^{m_A} \sum_{j=1}^{m_B} \left[u_{k-1;i,j}^{(N)}(t) - u_{k;i,j}^{(N)}(t) \right] \right\}^{m-m_1} \left\{ \sum_{i=1}^{m_A} \sum_{j=1}^{m_B} \left[u_{k;i,j}^{(N)}(t) \right] \right\}^{d-m} \\
&= \sum_{m=2}^d C_d^m \sum_{m_1-1=0}^{m-1} C_{m-1}^{m_1-1} \left\{ \sum_{j=1}^{m_B} \left[u_{k-1;l,j}^{(N)}(t) - u_{k;l,j}^{(N)}(t) \right] \right\}^{m_1-1} \\
&\times \left\{ \sum_{i \neq l}^{m_A} \sum_{j=1}^{m_B} \left[u_{k-1;i,j}^{(N)}(t) - u_{k;i,j}^{(N)}(t) \right] \right\}^{m-1-(m_1-1)} \left\{ \sum_{i=1}^{m_A} \sum_{j=1}^{m_B} \left[u_{k;i,j}^{(N)}(t) \right] \right\}^{d-m} \\
&= \sum_{m=2}^d C_d^m \left\{ \sum_{i=1}^{m_A} \sum_{j=1}^{m_B} \left[u_{k-1;i,j}^{(N)}(t) - u_{k;i,j}^{(N)}(t) \right] \right\}^{m-1} \left\{ \sum_{i=1}^{m_A} \sum_{j=1}^{m_B} \left[u_{k;i,j}^{(N)}(t) \right] \right\}^{d-m},
\end{aligned}$$

we have

$$\begin{aligned}
&C_d^1 \left\{ \sum_{i=1}^{m_A} \sum_{j=1}^{m_B} \left[u_{k;l,j}^{(N)}(t) \right] \right\}^{d-1} \\
&+ \sum_{m=2}^d C_d^m \left\{ \sum_{i=1}^{m_A} \sum_{j=1}^{m_B} \left[u_{k-1;i,j}^{(N)}(t) - u_{k;i,j}^{(N)}(t) \right] \right\}^{m-1} \left\{ \sum_{i=1}^{m_A} \sum_{j=1}^{m_B} \left[u_{k;i,j}^{(N)}(t) \right] \right\}^{d-m} \\
&= \sum_{m=1}^d C_d^m \left\{ \sum_{i=1}^{m_A} \sum_{j=1}^{m_B} \left[u_{k-1;i,j}^{(N)}(t) - u_{k;i,j}^{(N)}(t) \right] \right\}^{m-1} \left\{ \sum_{i=1}^{m_A} \sum_{j=1}^{m_B} \left[u_{k;i,j}^{(N)}(t) \right] \right\}^{d-m}.
\end{aligned}$$

Thus for $k \geq 1$ we obtain

$$L_{k;l}^{(N)} \left(u_{k-1}^{(N)}(t), u_k^{(N)}(t) \right) = \sum_{m=1}^d C_d^m \left\{ \sum_{i=1}^{m_A} \sum_{j=1}^{m_B} \left[u_{k-1;i,j}^{(N)}(t) - u_{k;i,j}^{(N)}(t) \right] \right\}^{m-1} \\ \times \left\{ \sum_{i=1}^{m_A} \sum_{j=1}^{m_B} \left[u_{k;i,j}^{(N)}(t) \right] \right\}^{d-m},$$

which is independent of phase $l \in \{1, 2, \dots, m_A\}$. Thus we have

$$L_k^{(N)} \left(u_{k-1}^{(N)}(t), u_k^{(N)}(t) \right) = L_{k;l}^{(N)} \left(u_{k-1}^{(N)}(t), u_k^{(N)}(t) \right).$$

Similarly, for phase $l \in \{1, 2, \dots, m_A\}$, we have

$$L_{1;l}^{(N)} \left(\left[u_0^{(N)}(t) \otimes \alpha \right], u_1^{(N)}(t) \right) = \sum_{m=1}^d C_d^m \left[\sum_{i=1}^{m_A} \sum_{j=1}^{m_B} \left(u_{0;i}^{(N)}(t) \alpha_j - u_{1;i,j}^{(N)}(t) \right) \right]^{m-1} \\ \times \left[\sum_{i=1}^{m_A} \sum_{j=1}^{m_B} u_{1;i,j}^{(N)}(t) \right]^{d-m}.$$

This gives

$$L_1^{(N)} \left(\left[u_0^{(N)}(t) \otimes \alpha \right], u_1^{(N)}(t) \right) = L_{1;l}^{(N)} \left(\left[u_0^{(N)}(t) \otimes \alpha \right], u_1^{(N)}(t) \right)$$

This completes the proof. ■

Appendix B: The Mean-Field Limit

In this appendix, we use the operator semigroup to provide a mean-field limit for the sequence of Markov processes $\{\mathbf{U}^{(N)}(t), t \geq 0\}$, which indicates the asymptotic independence of the block-structured queueing processes in this supermarket model. Note that the limits of the sequences of Markov processes can usually be discussed by the three main techniques: Operator semigroups, martingales, and stochastic equations. Readers may refer to Ethier and Kurtz [4] for more details.

To use the operator semigroups of Markov processes, we first need to introduce some state spaces as follows. For the vectors $\mathbf{u}^{(N)} = \left(u_0^{(N)}, u_1^{(N)}, u_2^{(N)}(t), \dots \right)$ where $u_0^{(N)}$ is a probability vector of size m_A and the size of the row vector $u_k^{(N)}$ is $m_A m_B$ for $k \geq 1$, we

write

$$\begin{aligned}\tilde{\Omega}_N = & \left\{ \mathbf{u}^{(N)} : u_1^{(N)} \geq u_2^{(N)} \geq u_3^{(N)} \geq \cdots \geq 0, \right. \\ & 1 = u_0^{(N)} e \geq u_1^{(N)} e \geq u_2^{(N)} e \geq \cdots \geq 0, \\ & \left. Nu_k^{(N)} \text{ is a vector of nonnegative integers for } k \geq 0 \right\}.\end{aligned}$$

and

$$\Omega_N = \left\{ \mathbf{u}^{(N)} \in \tilde{\Omega}_N : \mathbf{u}^{(N)} e < +\infty \right\}.$$

At the same time, for the vector $\mathbf{u} = (u_0, u_1, u_2, \dots)$ where u_0 is a probability vector of size m_A and the size of the row vector u_k is $m_A m_B$ for $k \geq 1$, we set

$$\tilde{\Omega} = \{ \mathbf{u} : u_1 \geq u_2 \geq u_3 \geq \cdots \geq 0; \quad 1 = u_0 e \geq u_1 e \geq u_2 e \geq \cdots \geq 0 \}$$

and

$$\Omega = \left\{ \mathbf{u} \in \tilde{\Omega} : \mathbf{u} e < +\infty \right\}.$$

Obviously, $\Omega_N \subsetneq \Omega \subsetneq \tilde{\Omega}$ and $\Omega_N \subsetneq \tilde{\Omega}_N \subsetneq \tilde{\Omega}$.

In the vector space $\tilde{\Omega}$, we take a metric

$$\begin{aligned}\rho(\mathbf{u}, \mathbf{u}') = & \max \left\{ \max_{1 \leq i \leq m_A} \{ |u_{0;i} - u'_{0;i}| \}, \right. \\ & \left. \max_{\substack{0 \leq i \leq m_A \\ 0 \leq j \leq m_B}} \sup_{k \geq 1} \left\{ \frac{|u_{k;i,j} - u'_{k;i,j}|}{k+1} \right\} \right\}\end{aligned}\tag{67}$$

for $\mathbf{u}, \mathbf{u}' \in \tilde{\Omega}$. Note that under the metric $\rho(\mathbf{u}, \mathbf{u}')$, the vector space $\tilde{\Omega}$ is separable and compact.

B.1: The operator semigroup

For $\mathbf{g} \in \Omega_N$, we write

$$L_1(g_0 \otimes \alpha, g_1) = \sum_{m=1}^d C_d^m \left[\sum_{l=1}^{m_A} \sum_{j=1}^{m_B} (g_{0;l}(t) \alpha_j - g_{1;l,j}) \right]^{m-1} \left(\sum_{l=1}^{m_A} \sum_{j=1}^{m_B} g_{1;l,j} \right)^{d-m},$$

and for $k \geq 2$

$$L_k(g_{k-1}, g_k) = \sum_{m=1}^d C_d^m \left[\sum_{l=1}^{m_A} \sum_{j=1}^{m_B} (g_{k-1;l,j} - g_{k;l,j}) \right]^{m-1} \left[\sum_{l=1}^{m_A} \sum_{j=1}^{m_B} g_{k;l,j} \right]^{d-m}.$$

Now, we consider the infinite-dimensional Markov process $\{\mathbf{U}^{(N)}(t), t \geq 0\}$ on state space Ω_N (or $\tilde{\Omega}_N$ in a similar analysis) for $N = 1, 2, 3, \dots$. Note that the stochastic evolution of this supermarket model of N identical servers is described as the Markov process $\{\mathbf{U}^{(N)}(t), t \geq 0\}$, where

$$\frac{d}{dt} \left(\mathbf{U}^{(N)}(t) \right) = \mathbf{A}_N f \left(\mathbf{U}^{(N)}(t) \right),$$

where \mathbf{A}_N acting on functions $f : \Omega_N \rightarrow \mathbf{C}^1$ is the generating operator of the Markov process $\{\mathbf{U}^{(N)}(t), t \geq 0\}$,

$$\mathbf{A}_N = \mathbf{A}_N^{\text{A-In}} + \mathbf{A}_N^{\text{A-Transition}} + \mathbf{A}_N^{\text{S-Transition}} + \mathbf{A}_N^{\text{S-Out}}, \quad (68)$$

for $\mathbf{g} \in \Omega_N$

$$\begin{aligned} \mathbf{A}_N^{\text{A-In}} f(\mathbf{g}) = & N \sum_{k=2}^{\infty} \sum_{j=1}^{m_B} \sum_{i=1}^{m_A} \left[\sum_{l=1}^{m_A} \left(g_{k-1;l,j} d_{l,i} - g_{k;i,j} \sum_{q=1}^{m_A} d_{i,q} \right) L_k(g_{k-1}, g_k) \right] \\ & \times \left[f \left(\mathbf{g} + \frac{e_{k;i,j}}{N} \right) - f(\mathbf{g}) \right] \\ & + N \sum_{j=1}^{m_B} \sum_{i=1}^{m_A} \left[\sum_{l=1}^{m_A} \left(g_{0;l} d_{l,i} \alpha_j - g_{1;i,j} \sum_{q=1}^{m_A} d_{i,q} \right) L_1(g_0 \otimes \alpha, g_1) \right] \\ & \times \left[f \left(\mathbf{g} + \frac{e_{1;i,j}}{N} \right) - f(\mathbf{g}) \right], \end{aligned} \quad (69)$$

$$\begin{aligned} \mathbf{A}_N^{\text{A-Transition}} = & N \sum_{k=1}^{\infty} \sum_{j=1}^{m_B} \sum_{i=1}^{m_A} \sum_{l=1}^{m_A} \left[g_{k;l,j} c_{l,i} + g_{k;i,j} \sum_{q=1}^{m_A} d_{i,q} \right] \\ & \times \left[f \left(\mathbf{g} - \frac{e_{k;l,j}}{N} + \frac{e_{k;i,j}}{N} \right) - f(\mathbf{g}) \right] \\ & + N \sum_{i=1}^{m_A} \sum_{l=1}^{m_A} \left[g_{0;l} c_{l,i} + g_{0,i} \sum_{q=1}^{m_A} d_{i,q} \right] \\ & \times \left[f \left(\mathbf{g} - \frac{e_{0;l}}{N} + \frac{e_{0;i}}{N} \right) - f(\mathbf{g}) \right], \end{aligned} \quad (70)$$

$$\begin{aligned} \mathbf{A}_N^{\text{S-Transition}} = & N \sum_{k=1}^{\infty} \sum_{i=1}^{m_A} \sum_{j=1}^{m_B} \sum_{r=1}^{m_B} (g_{k;i,r} t_{r,j}) \\ & \times \left[f \left(\mathbf{g} - \frac{e_{k;i,r}}{N} + \frac{e_{k;i,j}}{N} \right) - f(\mathbf{g}) \right] \end{aligned} \quad (71)$$

and

$$\mathbf{A}_N^{\text{S-Out}} = N \sum_{k=1}^{\infty} \sum_{i=1}^{m_A} \sum_{j=1}^{m_B} \sum_{r=1}^{m_B} (g_{k+1;i,r} t_r^0 \alpha_j) \left[f(\mathbf{g}) - f \left(\mathbf{g} - \frac{e_{k;i,j}}{N} \right) \right], \quad (72)$$

where $\mathbf{e}_{k;l,j}$ is a row vector of infinite size with the $(k; i, j)$ th entry being one and all others being zero. Thus it follows from Equations (68) to (72) that

$$\begin{aligned}
\mathbf{A}_N f(\mathbf{g}) = & N \sum_{k=2}^{\infty} \sum_{j=1}^{m_B} \sum_{i=1}^{m_A} \left[\sum_{l=1}^{m_A} \left(g_{k-1;l,j} d_{l,i} - g_{k;i,j} \sum_{q=1}^{m_A} d_{i,q} \right) L_k(g_{k-1}, g_k) \right] \\
& \times \left[f\left(\mathbf{g} + \frac{\mathbf{e}_{k;i,j}}{N}\right) - f(\mathbf{g}) \right] \\
& + N \sum_{j=1}^{m_B} \sum_{i=1}^{m_A} \left[\sum_{l=1}^{m_A} \left(g_{0;l} d_{l,i} \alpha_j - g_{1;i,j} \sum_{q=1}^{m_A} d_{i,q} \right) L_1(g_1 \otimes \alpha, g_1) \right] \\
& \times \left[f\left(\mathbf{g} + \frac{\mathbf{e}_{1;i,j}}{N}\right) - f(\mathbf{g}) \right] \\
& + N \sum_{k=1}^{\infty} \sum_{j=1}^{m_B} \sum_{i=1}^{m_A} \sum_{l=1}^{m_A} \left(g_{k;l,j} c_{l,i} + g_{k;i,j} \sum_{q=1}^{m_A} d_{i,q} \right) \\
& \times \left[f\left(\mathbf{g} - \frac{\mathbf{e}_{k;l,j}}{N} + \frac{\mathbf{e}_{k;i,j}}{N}\right) - f(\mathbf{g}) \right] \\
& + N \sum_{i=1}^{m_A} \sum_{l=1}^{m_A} \left(g_{0;l} c_{l,i} + \sum_{q=1}^{m_A} d_{i,q} \right) \left[f\left(\mathbf{g} - \frac{\mathbf{e}_{0;l}}{N} + \frac{\mathbf{e}_{0;i}}{N}\right) - f(\mathbf{g}) \right] \\
& + N \sum_{k=1}^{\infty} \sum_{i=1}^{m_A} \sum_{j=1}^{m_B} \sum_{r=1}^{m_B} \left\{ (g_{k;i,r} t_{r,j}) \left[f\left(\mathbf{g} - \frac{\mathbf{e}_{k;i,r}}{N} + \frac{\mathbf{e}_{k;i,j}}{N}\right) - f(\mathbf{g}) \right] \right. \\
& \left. + (g_{k+1;i,r} t_r^0 \alpha_j) \left[f(\mathbf{g}) - f\left(\mathbf{g} - \frac{\mathbf{e}_{k;i,j}}{N}\right) \right] \right\}. \tag{73}
\end{aligned}$$

Remark 5 If the MAP is a Poisson process, then $m_A = 1$ and $C = -\lambda$ and $D = \lambda$; and if the PH service time distribution is exponential, then $m_B = 1$, $T = -\mu$ and $T^0 \alpha = \mu$. In this case, it is easy to check from (73) that

$$\begin{aligned}
\mathbf{A}_N f(\mathbf{g}) = & \lambda N \left(1 - g_1^d \right) \left[f\left(\mathbf{g} + \frac{\mathbf{e}_1}{N}\right) - f(\mathbf{g}) \right] \\
& + \lambda N \sum_{k=2}^{\infty} \left(g_{k-1}^d - g_k^d \right) \left[f\left(\mathbf{g} + \frac{\mathbf{e}_k}{N}\right) - f(\mathbf{g}) \right] \\
& - \mu N \sum_{n=1}^{\infty} (g_n - g_{n+1}) \left[f(\mathbf{g}) - f\left(\mathbf{g} - \frac{\mathbf{e}_n}{N}\right) \right],
\end{aligned}$$

which is the same as (1.5) for $d = 2$ in Vvedenskaya et al [32].

B.2: The mean-Field limit

We compute

$$\lim_{N \rightarrow \infty} \frac{f\left(\mathbf{g} + \frac{e_{k;i,j}}{N}\right) - f(\mathbf{g})}{\frac{1}{N}} = \frac{\partial}{\partial g_{k;i,j}} f(\mathbf{g}),$$

$$\lim_{N \rightarrow \infty} \frac{f(\mathbf{g}) - f\left(\mathbf{g} - \frac{e_{k;i,j}}{N}\right)}{\frac{1}{N}} = \frac{\partial}{\partial g_{k;i,j}} f(\mathbf{g})$$

and

$$\lim_{N \rightarrow \infty} \frac{f\left(\mathbf{g} - \frac{e_{k;l,j}}{N} + \frac{e_{k;i,j}}{N}\right) - f(\mathbf{g})}{\frac{1}{N}} = \frac{\partial}{\partial g_{k;i,j}} f(\mathbf{g}) - \frac{\partial}{\partial g_{k;l,j}} f(\mathbf{g}).$$

The operator semigroup of the Markov process $\{\mathbf{U}^{(N)}(t), t \geq 0\}$ is defined as $\mathbf{T}_N(t)$, where if $f : \Omega_N \rightarrow \mathbf{C}^1$, then for $\mathbf{g} \in \Omega_N$ and $t \geq 0$

$$\mathbf{T}_N(t)f(\mathbf{g}) = E[f(\mathbf{U}_N(t)) \mid \mathbf{U}_N(0) = \mathbf{g}]. \quad (74)$$

Note that \mathbf{A}_N is the generating operator of the operator semigroup $\mathbf{T}_N(t)$, it is easy to see that $\mathbf{T}_N(t) = \exp\{\mathbf{A}_N t\}$ for $t \geq 0$.

Definition 1 *A operator semigroup $\{\mathbf{S}(t) : t \geq 0\}$ on the Banach space $L = C(\tilde{\Omega})$ is said to be strongly continuous if $\lim_{t \rightarrow 0} \mathbf{S}(t)f = f$ for every $f \in L$; it is said to be a contractive semigroup if $\|\mathbf{S}(t)\| \leq 1$ for $t \geq 0$.*

Let $L = C(\tilde{\Omega})$ be the Banach space of continuous functions $f : \tilde{\Omega} \rightarrow \mathbf{R}$ with uniform metric $\|f\| = \max_{u \in \tilde{\Omega}} |f(u)|$, and similarly, let $L_N = C(\Omega_N)$. The inclusion $\Omega_N \subset \tilde{\Omega}$ induces a contraction mapping $\Pi_N : L \rightarrow L_N$, $\Pi_N f(u) = f(u)$ for $f \in L$ and $u \in \Omega_N$.

Now, we consider the limiting behavior of the sequence $\{(\mathbf{U}^{(N)}(t), t \geq 0\}$ of Markov processes for $N = 1, 2, 3, \dots$. Two formal limits for the sequence $\{\mathbf{A}_N\}$ of generating operators and for the sequence $\{\mathbf{T}_N(t)\}$ of semigroups are expressed as $\mathbf{A} = \lim_{N \rightarrow \infty} \mathbf{A}_N$

and $\mathbf{T}(t) = \lim_{N \rightarrow \infty} \mathbf{T}_N(t)$ for $t \geq 0$, respectively. It follows from (73) that as $N \rightarrow \infty$

$$\begin{aligned}
\mathbf{A}f(\mathbf{g}) = & \sum_{k=2}^{\infty} \sum_{j=1}^{m_B} \sum_{i=1}^{m_A} \left[\sum_{l=1}^{m_A} \left(g_{k-1;l,j} d_{l,i} - g_{k;i,j} \sum_{q=1}^{m_A} d_{i,q} \right) L_k(g_{k-1}, g_k) \right] \frac{\partial}{\partial g_{k;i,j}} f(\mathbf{g}) \\
& + \sum_{j=1}^{m_B} \sum_{i=1}^{m_A} \left[\sum_{l=1}^{m_A} \left(g_{0;l,i} \alpha_j - g_{1;i,j} \sum_{q=1}^{m_A} d_{i,q} \right) L_1(g_0 \otimes \alpha, g_1) \right] \frac{\partial}{\partial g_{1;i,j}} f(\mathbf{g}) \\
& + \sum_{k=1}^{\infty} \sum_{j=1}^{m_B} \sum_{i=1}^{m_A} \sum_{l=1}^{m_A} \left(g_{k;l,j} c_{l,i} + g_{k;i,j} \sum_{q=1}^{m_A} d_{i,q} \right) \left[\frac{\partial}{\partial g_{k;i,j}} f(\mathbf{g}) - \frac{\partial}{\partial g_{k;l,j}} f(\mathbf{g}) \right] \\
& + \sum_{i=1}^{m_A} \sum_{l=1}^{m_A} \left(g_{0;l,i} c_{l,i} + g_{0,i} \sum_{q=1}^{m_A} d_{i,q} \right) \left[\frac{\partial}{\partial g_{0;i}} f(\mathbf{g}) - \frac{\partial}{\partial g_{0;l}} f(\mathbf{g}) \right] \\
& + \sum_{k=1}^{\infty} \sum_{i=1}^{m_A} \sum_{j=1}^{m_B} \sum_{r=1}^{m_B} \left\{ (g_{k;i,r} t_{r,j}) \left[\frac{\partial}{\partial g_{k;i,j}} f(\mathbf{g}) - \frac{\partial}{\partial g_{k;r,r}} f(\mathbf{g}) \right] \right. \\
& \left. + (g_{k+1;i,r} t_r^0 \alpha_j) \frac{\partial}{\partial g_{k;i,j}} f(\mathbf{g}) \right\}. \tag{75}
\end{aligned}$$

We define a mapping: $\mathbf{g} \rightarrow \mathbf{u}(t, \mathbf{g})$, where $\mathbf{u}(t, \mathbf{g})$ is a solution to the system of differential vector equations (22) to (26). Note that the operator semigroup $\mathbf{T}(t)$ acts in the space L , thus if $f \in L$ and $\mathbf{g} \in \tilde{\Omega}$, then

$$\mathbf{T}(t)f(\mathbf{g}) = f(\mathbf{u}(t, \mathbf{g})). \tag{76}$$

From (73) and (75), it is easy to see that the operator semigroups $\mathbf{T}_N(t)$ and $\mathbf{T}(t)$ are strongly continuous and contractive, see, for example, Section 1.1 in Chapter one of Ethier and Kurtz [4]. We denote by $\mathcal{D}(\mathbf{A})$ the domain of the generating operator \mathbf{A} . It follows from (76) that if f is a function from L and has the partial derivatives $\frac{\partial}{\partial g_{k;i,j}} f(\mathbf{g}) \in L$ for $k \geq 1, 1 \leq i \leq m_A, 1 \leq j \leq m_B$, and $\sup_{k \geq 1, 1 \leq i \leq m_A, 1 \leq j \leq m_B} \left\{ \left| \frac{\partial}{\partial g_{k;i,j}} f(\mathbf{g}) \right| \right\} < \infty$, then $f \in \mathcal{D}(\mathbf{A})$.

Let \mathbf{D} be the set of all functions $f \in L$ that have the partial derivatives $\frac{\partial}{\partial g_{k;i,j}} f(\mathbf{g})$ and $\frac{\partial^2}{\partial g_{k_1;m,n} \partial g_{k_2;r,s}} f(\mathbf{g})$, and there exists $C = C(f) < +\infty$ such that

$$\sup_{\substack{k \geq 1 \\ 1 \leq i \leq m_A, 1 \leq j \leq m_B \\ \mathbf{g} \in \tilde{\Omega}}} \left\{ \left| \frac{\partial}{\partial g_{k;i,j}} f(\mathbf{g}) \right| \right\} < C \tag{77}$$

and

$$\sup_{\substack{k_1, k_2 \geq 1 \\ 1 \leq m, r \leq m_A, 1 \leq n, s \leq m_B \\ \mathbf{g} \in \tilde{\Omega}}} \left\{ \left| \frac{\partial^2}{\partial g_{k_1; m, n} \partial g_{k_2; r, s}} f(\mathbf{g}) \right| \right\} < C. \quad (78)$$

We call that $f \in L$ depends only on the first K subvectors if for $\mathbf{g}^{(1)}, \mathbf{g}^{(2)} \in \tilde{\Omega}$, it follows from $g_i^{(1)} = g_i^{(2)}$ for $1 \leq i \leq K$ that $f(\mathbf{g}^{(1)}) = f(\mathbf{g}^{(2)})$, where $g_i^{(1)}$ and $g_i^{(2)}$ are row vectors of size $m_A m_B$ for $1 \leq i \leq K$. A similar and simple proof of that in Proposition 2 in Vvedenskaya et al [32] can show that the set of functions from L that depends on the first finite subvectors is dense in L .

The following lemma comes from Proposition 1 in Vvedenskaya et al [32]. We restated it here for convenience of description.

Lemma 1 *Consider an infinite-dimensional system of differential equations: For $k \geq 0$,*

$$z_k(0) = c_k$$

and

$$\frac{dz_k(t)}{dt} = \sum_{i=0}^{\infty} z_i(t) a_{i,k}(t) + b_k(t),$$

and let $\sum_{i=0}^{\infty} |a_{i,k}(t)| \leq a$, $|b_k(t)| \leq b_0 \exp \{bt\}$, $|c_k| \leq \varrho$, $b_0 \geq 0$ and $a < b$. Then

$$z_k(t) \leq \varrho \exp \{at\} + \frac{b_0}{b-a} [\exp \{bt\} - \exp \{at\}].$$

Definition 2 *Let A be a closed linear operator on the Banach space $L = C(\tilde{\Omega})$. A subspace \mathbf{D} of $\mathcal{D}(A)$ is said to be a core for A if the closure of the restriction of A to \mathbf{D} is equal to A , i.e., $\overline{A|_{\mathbf{D}}} = A$.*

For any matrix $\mathbf{A} = (a_{i,j})$, we define its norm as follows:

$$\|\mathbf{A}\| = \max_i \left\{ \sum_j |a_{i,j}| \right\}.$$

It is easy to compute that

$$\|I \otimes \mathbf{A}\| = \|\mathbf{A}\|,$$

$$\|\mathbf{A} \otimes I\| = \|\mathbf{A}\|,$$

$$\|\text{diag}(De)\| = \|D\|.$$

We introduce some notation

$$M_1 = \sum_{m=1}^d C_d^m = 2^d - 1,$$

$$M_2 = m_A m_B \sum_{m=1}^d C_d^m (d + m - 2),$$

$$a = \|T^0 \alpha\| + \|[C + \text{diag}(De)] \oplus T\| + 2\|D\|(M_1 + M_2).$$

The following lemma is a key to prove that the set \mathbf{D} is a core for the generating operator \mathbf{A} .

Lemma 2 *Let $\mathbf{u}(t)$ be a solution to the system of differential vector equations (22) to (23). Then*

$$\sup_{\substack{k, k_1 \geq 1 \\ 1 \leq i, m \leq m_A, 1 \leq j, n \leq m_B}} \left\{ \left| \frac{\partial u_{k;i,j}(t, \mathbf{g})}{\partial g_{k_1;m,n}} \right| \right\} \leq \varrho \exp \{at\}, \quad (79)$$

and

$$\sup_{\substack{k, k_1, k_2 \geq 1 \\ 1 \leq i, m, r \leq m_A \\ 1 \leq j, n, s \leq m_B}} \left\{ \left| \frac{\partial^2 u_{k;i,j}(t, \mathbf{g})}{\partial g_{k_1;m,n} \partial g_{k_2;r,s}} \right| \right\} \leq \widehat{\varrho} \exp \{at\} + \frac{2\|D\|}{a} (\exp \{2at\} - \exp \{at\}). \quad (80)$$

Proof: We only prove Inequalities (79), while Inequalities (80) can be proved similarly.

Notice that $\mathbf{u}(t)$ is a solution to the system of differential vector equations (22) to (23) and possesses the derivatives $\frac{\partial u_{k;i,j}(t, \mathbf{g})}{\partial g_{k_1;m,n}}$ and $\frac{\partial^2 u_{k;i,j}(t, \mathbf{g})}{\partial g_{k_1;m,n} \partial g_{k_2;r,s}}$. For simplicity of description, we set $u'_{k;i,j,k_1;m,n} = \frac{\partial u_{k;i,j}(t, \mathbf{g})}{\partial g_{k_1;m,n}}$. It follows from (22) to (23) that for $k, k_1 \geq 1, 1 \leq i, m \leq m_A$ and $1 \leq j, n \leq m_B$,

$$\begin{aligned} \frac{du'_{k;i,j,k_1;m,n}}{dt} = & \sum_{l=1}^{m_A} \left(u'_{k-1;l,j,k_1;m,n} d_{l,i} - u'_{k;i,j,k_1;m,n} \sum_{q=1}^{m_A} d_{i,q} \right) L_k(u_{k-1}(t), u_k(t)) \\ & + \sum_{l=1}^{m_A} \left(u_{k-1;l,j} d_{l,i} - u_{k;l,j} \sum_{q=1}^{m_A} d_{i,q} \right) L'_k(u_{k-1}(t), u_k(t)) \\ & + \sum_{l=1}^{m_A} u'_{k;l,j,k_1;m,n} c_{l,i} + u'_{k;i,j,k_1;m,n} \sum_{q=1}^{m_A} d_{i,q} \\ & + \sum_{s=1}^{m_B} u'_{k;i,s,k_1;m,n} t_{s,j} + \sum_{s=1}^{m_B} u'_{k+1;i,s,k_1;m,n} t_s^0 \alpha_j, \end{aligned}$$

and

$$\begin{aligned}
L'_k(u_{k-1}(t), u_k(t)) &= \sum_{m=1}^d C_d^m(m-1) \left[\sum_{l=1}^{m_A} \sum_{j=1}^{m_B} (u_{k-1;l,j} - u_{k;l,j}) \right]^{m-2} \\
&\times \left[\sum_{l=1}^{m_A} \sum_{j=1}^{m_B} u_{k;l,j} \right]^{d-m} \left[\sum_{l=1}^{m_A} \sum_{j=1}^{m_B} (u'_{k-1;l,j,k_1;m,n} - u'_{k;l,j,k_1;m,n}) \right] \\
&+ \sum_{m=1}^d C_d^m(d-m) \left[\sum_{l=1}^{m_A} \sum_{j=1}^{m_B} (u_{k-1;l,j} - u_{k;l,j}) \right]^{m-1} \\
&\times \left[\sum_{l=1}^{m_A} \sum_{j=1}^{m_B} u_{k;l,j} \right]^{d-m-1} \left[\sum_{l=1}^{m_A} \sum_{j=1}^{m_B} u'_{k;l,j,k_1;m,n} \right].
\end{aligned}$$

Using Lemma 1, we obtain Inequalities (79) with

$$\begin{aligned}
a &= \|I \otimes T^0 \alpha\| + \|[C + \text{diag}(De)] \oplus T\| \\
&\quad + [\|D \otimes I\| + \|\text{diag}(De) \otimes I\|] (M_1 + M_2) \\
&= \|T^0 \alpha\| + \|[C + \text{diag}(De)] \oplus T\| + 2\|D\| (M_1 + M_2),
\end{aligned}$$

$$b_0 = 0$$

and

$$\varrho = \sup_{\substack{k, k_1 \geq 1 \\ 1 \leq i, m \leq m_A, 1 \leq j, n \leq m_B}} \{ |u'_{k;i,j,k_1;m,n}(0)| \}.$$

This completes this proof. ■

Lemma 3 *The set \mathbf{D} is a core for the operator A .*

Proof: It is obvious that \mathbf{D} is dense in L and $\mathbf{D} \in \mathcal{D}(A)$. Let \mathbf{D}_0 be the set of functions from \mathbf{D} which depend only on the first K subvectors of size $m_A m_B$. It is easy to see that \mathbf{D}_0 is dense in L . Therefore, Using proposition 3.3 in Chapter 1 of Ethier and Kurtz [4], it can show that for any $t \geq 0$, the operator $\mathbf{T}(t)$ does not bring \mathbf{D}_0 out of \mathbf{D} . Select an arbitrary function $\varphi \in \mathbf{D}_0$ and let $f(\mathbf{g}) = \varphi(\mathbf{u}(t; \mathbf{g}))$, $\mathbf{g} \in \tilde{\Omega}$. It follows from Lemma 2 that f has partial derivatives $\frac{\partial}{\partial g_{k;i,j}} f(\mathbf{g})$ and $\frac{\partial^2}{\partial g_{k_1;m,n} \partial g_{k_2;r,s}} f(\mathbf{g})$ that satisfy conditions (77) and (78). Therefore $f \in \mathbf{D}$. This completes the proof. ■

In what follows we can prove Theorem 2 given in Section 5.

Proof of Theorem 2: This proof is to use the convergence of operator semigroups as well as the convergence of their corresponding generating generators, e.g., see Theorem

6.1 in Chapter 1 of Ethier and Kurtz [4]. Lemma 3 shows that the set \mathbf{D} is a core for the generating operator \mathbf{A} . For any function $f \in \mathbf{D}$, we have

$$N \left[f \left(\mathbf{g} - \frac{e_{n;i,j}}{N} \right) - f(\mathbf{g}) \right] - \frac{\partial}{\partial g_{n;i,j}} f(\mathbf{g}) = -\frac{\gamma_{n;i,j}^{(1)}}{N} \frac{\partial^2 f \left(\mathbf{g} - \gamma_{n;i,j}^{(2)} \right)}{\partial g_{n;i,j}^2},$$

and

$$\left\| \frac{\gamma_{n;i,j}^{(1)}}{N} \frac{\partial^2 f \left(\mathbf{g} - \gamma_{n;i,j}^{(2)} \right)}{\partial g_{n;i,j}^2} \right\| \leq \frac{\Re}{N}.$$

Thus we obtain

$$\begin{aligned} |\mathbf{A}_N f(\mathbf{g}) - f(\mathbf{g})| \leq & \frac{\Re}{N} \left\{ \sum_{k=2}^{\infty} \sum_{j=1}^{m_B} \sum_{i=1}^{m_A} \left[\sum_{l=1}^{m_A} \left(g_{k-1;l,j} d_{l,i} - g_{k;i,j} \sum_{q=1}^{m_A} d_{i,q} \right) L_k(g_{k-1}, g_k) \right] \right. \\ & + \sum_{j=1}^{m_B} \sum_{i=1}^{m_A} \left[\sum_{l=1}^{m_A} \left(g_{0;l} d_{l,i} \alpha_j - g_{1;i,j} \sum_{q=1}^{m_A} d_{i,q} \right) L_1(g_0 \otimes \alpha, g_1) \right] \\ & + \sum_{k=1}^{\infty} \sum_{j=1}^{m_B} \sum_{i=1}^{m_A} \sum_{l=1}^{m_A} \left(g_{k;l,j} |c_{l,i}| + g_{k;i,j} \sum_{q=1}^{m_A} d_{i,q} \right) \\ & + \sum_{i=1}^{m_A} \sum_{l=1}^{m_A} \left(g_{0;l} |c_{l,i}| + g_{0,i} \sum_{q=1}^{m_A} d_{i,q} \right) \\ & \left. + \sum_{k=1}^{\infty} \sum_{i=1}^{m_A} \sum_{j=1}^{m_B} \sum_{l=1}^{m_B} (g_{k;i,l} |t_{l,j}| + g_{k+1;i,l} t_l^0 \alpha_j) \right\}. \end{aligned}$$

Note that

$$L_1(g_0 \otimes \alpha, g_1) = \frac{(g_0 e)^d - (g_1 e)^d}{g_0 e - g_1 e} \leq d$$

and

$$L_1(g_{k-1}, g_k) = \frac{(g_{k-1} e)^d - (g_k e)^d}{g_{k-1} e - g_k e} \leq d,$$

we obtain

$$\begin{aligned} |\mathbf{A}_N f(\mathbf{g}) - f(\mathbf{g})| \leq & \frac{\Re}{N} \left[d \sum_{k=2}^{\infty} \sum_{j=1}^{m_B} \sum_{i=1}^{m_A} \sum_{l=1}^{m_A} g_{k-1;l,j} d_{l,i} + d \sum_{j=1}^{m_B} \sum_{i=1}^{m_A} \sum_{l=1}^{m_A} g_{0;l} d_{l,i} \alpha_j \right. \\ & \left. + (\|C\| + m_A \|D\| + \|T\| + \|T^0 \alpha\|) \sum_{k=0}^{\infty} g_k e \right] \\ \leq & \frac{\Re}{N} \left[(\|C\| + (d + m_A) \|D\| + \|T\| + \|T^0 \alpha\|) \sum_{k=0}^{\infty} g_k e \right]. \end{aligned}$$

For $\mathbf{g} \in \Omega$, it is clear that $\mathbf{g}e = \sum_{k=0}^{\infty} g_k e < +\infty$. Thus we get

$$\lim_{N \rightarrow \infty} \sup_{\mathbf{g} \in \Omega} |\mathbf{A}_N f(\mathbf{g}) - \mathbf{A} f(\mathbf{g})| = 0.$$

This gives

$$\lim_{N \rightarrow \infty} \sup_{\mathbf{g} \in \Omega} |\mathbf{T}_N(t) f(\mathbf{g}) - f(\mathbf{u}(t; \mathbf{g}))| = 0.$$

This completes the proof. ■

Appendix C: Proof of Theorem 3

To prove Theorem 3, we need to extend the coupling method given in Turner [30] and Martin and Suhov [22] such that this coupling method can be applied to discussing stability of more general block-structured supermarket models.

In the two supermarket models Q and R , they have the same parameters: $N, d, m_A, C, D, m_B, \alpha, T$, and the same initial state at $t = 0$; while the only difference between both of them is their choice numbers: $d(Q) = 1$ and $d(R) \geq 2$.

To set up a coupling between the two infinite-dimensional Markov processes $\{U_N^{(Q)}(t) : t \geq 0\}$ and $\{U_N^{(R)}(t) : t \geq 0\}$, we need introduce some notation as follows. For a supermarket model S with $k \geq 1$, $1 \leq i \leq m_A$ and $1 \leq j \leq m_B$, we denote by $A_k^{(i,j)}(S)$ and $D_k^{(i,j)}(S)$ the k th arrival time and the k th departure time when the MAP environment process is at state i and the PH service environment process is at state j .

As discussed in Section 4 of Martin and Suhov [22], we introduce the notation of "shadow" customers to build up the coupling relation between the two supermarket models Q and R . For k and (i, j) , the time of the shadow customer arriving at the supermarket model Q is written as $A_k^{(i,j)}(R)$, and at time $A_k^{(i,j)}(Q)$ the shadow customer is replaced by the real customer immediately. The relationship between the shadow and real customers are described by Figure 8 (a), while there will not exist a shadow customer in Figure 8 (b).

From the two supermarket models Q and R , we construct a new supermarket model \overline{Q} with shadow customers such that at environment state pair (i, j) , each arrival time in the supermarket model \overline{Q} is the same time as that in the supermarket model R , while each departure time is the same time as that in supermarket model Q . Based on this,

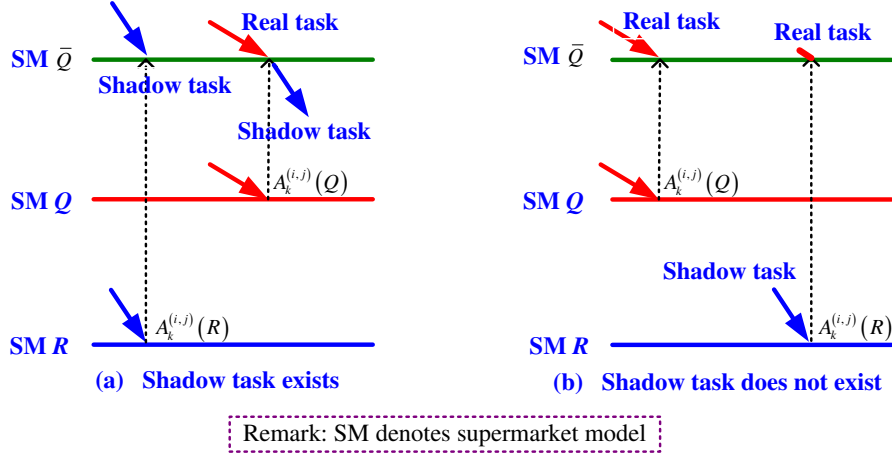


Figure 7: The shadow and real tasks

we can set up a coupling between the two supermarket models R and Q by means of the supermarket model \bar{Q} .

For a supermarket model S and for $k \geq 1, 1 \leq i \leq m_A, 1 \leq i \leq m_B, x \geq 0$, we define

$$\psi_x^{(i,j)}(S, t) = \sum_{n=1}^N \left[l_n^{(i,j)}(S, t) - x \right]_+,$$

where $l_n^{(i,j)}(S, t)$ is the queue length of the n th server with environment state pair (i, j) at time t , and $[y]_+ = \max(y, 0)$.

The following lemma gives a useful property of $\psi_x^{(i,j)}(S, t)$ for the two supermarket models \bar{Q} and R .

Lemma 4 *If $\psi_y^{(i,j)}(R, t) \leq \psi_y^{(i,j)}(\bar{Q}, t)$ for all y and $\psi_x^{(i,j)}(R, t) = \psi_x^{(i,j)}(\bar{Q}, t)$, then*

$$\# \left\{ n : l_n^{(i,j)}(R, t) \leq x \right\} \leq \# \left\{ n : l_n^{(i,j)}(\bar{Q}, t) \leq x \right\} \quad (81)$$

and

$$\# \left\{ n : l_n^{(i,j)}(R, t) \geq x \right\} \leq \# \left\{ n : l_n^{(i,j)}(\bar{Q}, t) \geq x \right\}, \quad (82)$$

where $\# \{A\}$ means the number of elements in the set A .

Proof: If $\psi_y^{(i,j)}(R, t) \leq \psi_y^{(i,j)}(\bar{Q}, t)$ for all y and $\psi_x^{(i,j)}(R, t) = \psi_x^{(i,j)}(\bar{Q}, t)$, then for $y = x + 1$

$$-\psi_{x+1}^{(i,j)}(R, t) \geq -\psi_{x+1}^{(i,j)}(\bar{Q}, t),$$

using $\psi_x^{(i,j)}(R, t) = \psi_x^{(i,j)}(\overline{Q}, t)$ we get

$$\psi_x^{(i,j)}(R, t) - \psi_{x+1}^{(i,j)}(R, t) \geq \psi_x^{(i,j)}(\overline{Q}, t) - \psi_{x+1}^{(i,j)}(\overline{Q}, t). \quad (83)$$

Similarly, for $y = x - 1$ we have

$$\psi_x^{(i,j)}(R, t) - \psi_{x-1}^{(i,j)}(R, t) \leq \psi_x^{(i,j)}(\overline{Q}, t) - \psi_{x-1}^{(i,j)}(\overline{Q}, t). \quad (84)$$

Since

$$\psi_x^{(i,j)}(S, t) = \sum_{n=1}^N \left[l_n^{(i,j)}(S, t) - x \right]_+,$$

we obtain

$$\psi_x^{(i,j)}(S, t) - \psi_{x+1}^{(i,j)}(S, t) = \sum_{n=1}^N \left\{ \left[l_n^{(i,j)}(S, t) - x \right]_+ - \left[l_n^{(i,j)}(S, t) - (x+1) \right]_+ \right\}.$$

To calculate $\psi_x^{(i,j)}(S, t) - \psi_{x+1}^{(i,j)}(S, t)$, we analyze the following two cases:

Case one: If $l_n^{(i,j)}(S, t) \leq x$, then $\left[l_n^{(i,j)}(S, t) - x \right]_+ = \left[l_n^{(i,j)}(S, t) - (x+1) \right]_+ = 0$.

Case two: If $l_n^{(i,j)}(S, t) > x$, then $\left[l_n^{(i,j)}(S, t) - x \right]_+ - \left[l_n^{(i,j)}(S, t) - (x+1) \right]_+ = 1$.

If $\sum_{n=1}^N \left\{ \left[l_n^{(i,j)}(S, t) - x \right]_+ - \left[l_n^{(i,j)}(S, t) - (x+1) \right]_+ \right\} = k$, then k is the number of servers whose queue length is bigger than x . That is $\# \left\{ n : l_n^{(i,j)}(S, t) > x \right\} = k$. Hence, we obtain

$$\begin{aligned} \psi_x^{(i,j)}(S, t) - \psi_{x+1}^{(i,j)}(S, t) &= \sum_{n=1}^N \left\{ \left[l_n^{(i,j)}(S, t) - x \right]_+ - \left[l_n^{(i,j)}(S, t) - (x+1) \right]_+ \right\} \\ &= \# \left\{ n : l_n^{(i,j)}(S, t) > x \right\}. \end{aligned} \quad (85)$$

It follows from (83) to (85) that

$$\# \left\{ n : l_n^{(i,j)}(R, t) > x \right\} \geq \# \left\{ n : l_n^{(i,j)}(\overline{Q}, t) > x \right\},$$

this gives

$$\# \left\{ n : l_n^{(i,j)}(R, t) \leq x \right\} \leq \# \left\{ n : l_n^{(i,j)}(\overline{Q}, t) \leq x \right\}.$$

Similarly, it follows from (84) to (85) that

$$\# \left\{ n : l_n^{(i,j)}(R, t) > x - 1 \right\} \leq \# \left\{ n : l_n^{(i,j)}(\overline{Q}, t) > x - 1 \right\},$$

which follows

$$\# \left\{ n : l_n^{(i,j)}(R, t) \geq x \right\} \leq \# \left\{ n : l_n^{(i,j)}(\overline{Q}, t) \geq x \right\}.$$

This completes the proof. ■

The following lemma sets up the coupling between the two supermarket models R and \overline{Q} , which is based on the arrival and departure processes.

Lemma 5 *For the two supermarket models R and \overline{Q} and for $x, t \geq 0, 1 \leq i \leq m_A, 1 \leq i \leq m_B$, we have*

$$\psi_x^{(i,j)}(R, t) \leq \psi_x^{(i,j)}(\overline{Q}, t). \quad (86)$$

Proof: To prove (86), we need to discuss the departure process and the arrival process, respectively.

(1) The departure process

Note that the two supermarket models R and \overline{Q} have the same initial state at $t = 0$, thus (86) holds at time $t = 0$.

In the departing process, it is easy to see from the above coupling that at environment state pair (i, j) , if given the server orders in supermarket models \overline{Q} and R according to the queue length of each server (including shadow tasks), then the customer departures always occur at the same order servers. For example, if the customer departure occurs from the server with the shortest queue length in supermarket model \overline{Q} , then a customer departure must also occur from the server with the shortest queue length in supermarket model R . Note that the customer departures will be lost either from an empty server or from one containing only shadow customers.

Let D be a potential departure time at environment state pair (i, j) , and suppose that (86) holds for $t < D$. Then we hope to show that (86) holds for $t = D$.

Suppose that (86) does not hold at a departure point D . Then we have $\psi_x^{(i,j)}(R, D) > \psi_x^{(i,j)}(\overline{Q}, D)$.

Since (86) holds for $t < D$, we get that $\psi_x^{(i,j)}(\overline{Q}, D^-) \leq \psi_x^{(i,j)}(R, D^-)$. Based on this, we discuss the two cases: $\psi_x^{(i,j)}(\overline{Q}, D^-) = \psi_x^{(i,j)}(R, D^-)$ and $\psi_x^{(i,j)}(\overline{Q}, D^-) < \psi_x^{(i,j)}(R, D^-)$, and indicate how the two cases influence the departure process at time D .

Case one: If $\psi_x^{(i,j)}(\overline{Q}, D^-) = \psi_x^{(i,j)}(R, D^-)$ and $\psi_x^{(i,j)}(R, D) > \psi_x^{(i,j)}(\overline{Q}, D)$, then a departure at time D makes that $\psi_x^{(i,j)}(R, D)$ does not change, while $\psi_x^{(i,j)}(\overline{Q}, D)$ is diminished. Let a and b be the queue lengths at time D in the two supermarket models

\overline{Q} and R , respectively. Then for $x = 0, 1, \dots, a-1$, it is seen that

$$\psi_x^{(i,j)}(\overline{Q}, D^-) = \sum_{n=1}^N \left[l_n^{(i,j)}(\overline{Q}, t) - x \right]_+$$

reduces 1. Similarly, for $x = 0, 1, \dots, b-1$,

$$\psi_x^{(i,j)}(R, D^-) = \sum_{n=1}^N \left[l_n^{(i,j)}(R, t) - x \right]_+$$

also reduce 1. Therefore, when x is $b, b+1, \dots, a-1$ (that is $b \leq x < a$), we have $\psi_x^{(i,j)}(R, D) > \psi_x^{(i,j)}(\overline{Q}, D)$. However, when $\psi_x^{(i,j)}(\overline{Q}, D^-) = \psi_x^{(i,j)}(R, D^-)$, both from that (81) holds for $t < D$ and from that the departure channels are at a coupling, it is clear that the condition: $\psi_x^{(i,j)}(R, D) > \psi_x^{(i,j)}(\overline{Q}, D)$ for $b \leq x < a$, is impossible.

Case two: $\psi_x^{(i,j)}(R, D^-) < \psi_x^{(i,j)}(\overline{Q}, D^-)$. In this case, when a customer departs the system, the two numbers $\psi_x^{(i,j)}(R, D^-)$ and $\psi_x^{(i,j)}(\overline{Q}, D^-)$ have only two cases: Unchange and diminish 1. Note that $\psi_x^{(i,j)}(R, D^-) < \psi_x^{(i,j)}(\overline{Q}, D^-)$, we get that $\psi_x^{(i,j)}(R, D^-) + 1 \leq \psi_x^{(i,j)}(\overline{Q}, D^-)$. Hence, we can not obtain that $\psi_x^{(i,j)}(R, D) > \psi_x^{(i,j)}(\overline{Q}, D)$.

(2) The arrival process

In a similar way to the above analysis in "(1) The departure process", we discuss the coupling for the arriving process as follows.

Let $A = A_k^{(i,j)}$ be an arrival time. Then (86) holds for $t < A$. We hope to show that (86) holds for $t = A$.

This proof is similar to the above analysis in "(1) The departure process". Let a and b be the queue lengths at time A in the two supermarket models \overline{Q} and R , respectively. Then $\psi_x^{(i,j)}(R, A^-) = \psi_x^{(i,j)}(\overline{Q}, A^-)$ holds for some x for $a < x \leq b$. Thus, it follows from (82) that

$$\# \left\{ n : l_n^{(i,j)}(R, A^-) \geq x \right\} \leq \# \left\{ n : l_n^{(i,j)}(\overline{Q}, A^-) \geq x \right\}$$

and

$$\# \left\{ n : l_n^{(i,j)}(R, A^-) \geq b \right\} \leq \# \left\{ n : l_n^{(i,j)}(\overline{Q}, A^-) \geq a \right\}.$$

However, the condition: $\# \left\{ n : l_n^{(i,j)}(R, A^-) \geq b \right\} \leq \# \left\{ n : l_n^{(i,j)}(\overline{Q}, A^-) \geq a \right\}$, is impossible, because it follows from the above coupling that for $a < x \leq b$

$$\# \left\{ n : l_n^{(i,j)}(R, A^-) \geq b \right\} > \# \left\{ n : l_n^{(i,j)}(\overline{Q}, A^-) \geq a \right\}.$$

Since the queue length a was chosen at the arrival time, it is seen that the queue length a must exist in the supermarket model \overline{Q} . In this case, we get that $\# \left\{ n : l_n^{(i,j)}(\overline{Q}, A^-) = a \right\} \geq 1$. Therefore, this leads to a contradiction.

Note that there are some shadow customers in supermarket model \overline{Q} , the shadow customers do not affect the queue lengths in the supermarket model \overline{Q} at the arrival time $A_k^{(i,j)}(Q)$, thus (86) holds. This completes the proof. ■

The following lemma provides the coupling between the two supermarket models Q and R , which is based on the arrival and departure processes.

Lemma 6 *In the two supermarket models Q and R , for $k > 0, 1 \leq i \leq m_A, 1 \leq j \leq m_B$ we have*

$$D_k^{(i,j)}(R) \leq D_k^{(i,j)}(Q) \quad (87)$$

and

$$A_k^{(i,j)}(R) \leq A_k^{(i,j)}(Q). \quad (88)$$

Proof: Using the above coupling, now we continue to discuss the two supermarket models Q and R .

Note that the two supermarket models Q and R have the same parameters $N, m, c_{i,j}, d_{i,j}, \mu_i$ for $1 \leq i, j \leq m$ and the same initial state at $t = 0$, the departure or arrival of the k th customer and the Markov environment process in the supermarket model Q correspond to those in the supermarket model R . This ensures that if (87) holds for the departure process up to a given time, then so does (88) for the arrival process up to that time.

Now, we use (86) to prove (87).

Suppose that (87) is false, that is, $D_k^{(i,j)}(R) > D_k^{(i,j)}(Q)$. Then the number of customer departures before time D from the supermarket model R must be the same as that in the supermarket model \overline{Q} . Since the arrivals in the two supermarket models R and \overline{Q} occur at the same times, there must be the same total number of customers in the two supermarket models R and \overline{Q} . Hence, $\psi_0^{(i,j)}(R, D^-) = \psi_0^{(i,j)}(\overline{Q}, D^-)$. But, it is seen from (81) that the number of servers with non-zero queue length in the supermarket model \overline{Q} is bigger than that in the supermarket model R , this indicates that the number of servers with empty server in the supermarket model \overline{Q} is less than that in the supermarket model R . Therefore, if a departure occurs in the supermarket model \overline{Q} , then there must be a departure in the supermarket model R . On the contrary, if a departure occurs in the

supermarket model R , then it is possible not to have a departure in the supermarket model \overline{Q} . Note that the departure time in the supermarket model \overline{Q} is the same as that in the supermarket model Q , hence the departure time in the supermarket model R is earlier than that in the supermarket model Q , that is, $D_k^{(i,j)}(R) \leq D_k^{(i,j)}(Q)$. This leads to a contradiction of the assumption $D_k^{(i,j)}(R) > D_k^{(i,j)}(Q)$. Hence (87) holds. Similarly, we can prove (88). This completes the proof. ■

Proof of Theorem 3: Using the lemma 6, we know that $D_k^{(i,j)}(R) \leq D_k^{(i,j)}(Q)$ and $A_k^{(i,j)}(R) \leq A_k^{(i,j)}(Q)$. This indicates that for any two corresponding servers in the two supermarket models Q and R , the arrival and departure times in the supermarket model R are earlier than those in the supermarket model Q . Hence, the queue length of any server in the supermarket model R is shorter than that of the corresponding server in the supermarket model Q . This shows that the total number of customers in the supermarket model R is no greater than the total number of customers in the supermarket model Q at time $t \geq 0$. Based on this, we obtain a coupling between the processes $\{U_Q^{(N)}(t)\}$ and $\{U_R^{(N)}(t)\}$: For all $t \geq 0$, the total number of customers in the supermarket model R is no greater than that in the supermarket model Q . This completes the proof. ■

References

- [1] Bramson M, Lu Y, Prabhakar B (2010) Randomized load balancing with general service time distributions. In: Proceedings of the ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems, pp 275–286
- [2] Bramson M, Lu Y, Prabhakar B (2012) Asymptotic independence of queues under randomized load balancing. Queueing Syst 71:247–292
- [3] Bramson M, Lu Y, Prabhakar B (2013) Decay of tails at equilibrium for FIFO join the shortest queue networks. Ann Appl Probab 23:1841–1878
- [4] Ethier SN, Kurtz TG (1986) Markov Processes: Characterization and Convergence. John Wiley & Sons, New York
- [5] Graham C (2000) Kinetic limits for large communication networks. In: N. Bellomo and M. Pulvirenti (eds.) Modelling in Applied Sciences. Birkhäuser, pp 317–370

- [6] Graham C (2000) Chaoticity on path space for a queueing network with selection of the shortest queue among several. *J Appl Probab* 37:198–201
- [7] Graham C (2004) Functional central limit theorems for a large network in which customers join the shortest of several queues. *Probab Theory Relat Fields* 131:97–120
- [8] Jacquet P, Vvedenskaya ND (1998) On/off sources in an interconnection networks: Performance analysis when packets are routed to the shortest queue of two randomly selected nodes. Technical Report N^o 3570, INRIA Rocquencourt, France
- [9] Jacquet P, Suhov YM, Vvedenskaya ND (1999) Dynamic routing in the mean-field approximation. Technical Report N^o 3789, INRIA Rocquencourt, France
- [10] Kurtz TG (1981) *Approximation of Population Processes*. SIAM
- [11] Li QL (2010) *Constructive Computation in Stochastic Models with Applications: The RG-Factorizations*. Springer and Tsinghua Press
- [12] Li QL (2011) Super-exponential solution in Markovian supermarket models: Framework and challenge. Available: [arXiv:1106.0787](https://arxiv.org/abs/1106.0787)
- [13] Li QL (2014) Tail probabilities in queueing processes. *Asia-Pacific Journal of Operational Research* 31:1–31 (No. 2)
- [14] Li QL, Cao J (2004) Two types of *RG*-factorizations of quasi-birth-and-death processes and their applications to stochastic integral functionals. *Stochastic Models* 20:299-340
- [15] Li QL, Dai G, Lui JCS, Wang Y (2013) The mean-field computation in a supermarket model with server multiple vacations. *Discrete Event Dyn Syst*, Available in Publishing Online: November 8, 2013, Pages 1–50
- [16] Li QL, Lui JCS (2010) Doubly exponential solution for randomized load balancing models with Markovian arrival processes and PH service times. Available: [arXiv:1105.4341](https://arxiv.org/abs/1105.4341)
- [17] Li QL, Lui JCS, Wang Y (2011) A matrix-analytic solution for randomized load balancing models with PH service times. In: *Performance Evaluation of Computer*

and Communication Systems: Milestones and Future Challenges. Lecture Notes in Computer Science, vol 6821, pp 240–253

- [18] Luczak MJ, McDiarmid C (2006) On the maximum queue length in the supermarket model. *Ann Probab* 34:493–527
- [19] Luczak MJ, McDiarmid C (2007) Asymptotic distributions and chaos for the supermarket model. *Electron J Probab* 12:75–99
- [20] Luczak MJ, Norris JR (2005) Strong approximation for the supermarket model. *Ann Appl Probab* 15:2038–2061
- [21] Martin JB (2001) Point processes in fast Jackson networks. *Ann Appl Probab* 11:650–663
- [22] Martin JB, Suhov YM (1999) Fast Jackson networks. *Ann Appl Probab* 9:854–870
- [23] Mitzenmacher MD (1996) The Power of Two Choices in Randomized Load Balancing. PhD Thesis, Department of Computer Science, University of California at Berkeley, USA
- [24] Mitzenmacher MD (1999) On the analysis of randomized load balancing schemes. *Theory Comput Syst* 32:361–386
- [25] Mitzenmacher MD, Richa A, Sitaraman R (2001) The power of two random choices: A survey of techniques and results. In: *Handbook of randomized computing*, vol 1, pp 255–312
- [26] Mitzenmacher MD, Upfal E (2005) *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press
- [27] Neuts MF (1981) *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach*. Johns Hopkins University Press
- [28] Neuts MF (1989) *Structured Stochastic Matrices of $M/G/1$ Type and Their Applications*. Marcel Dekker Inc., New York.
- [29] Suhov YM, Vvedenskaya ND (2002) Fast Jackson Networks with Dynamic Routing. *Probl Inf Transm* 38:136–153

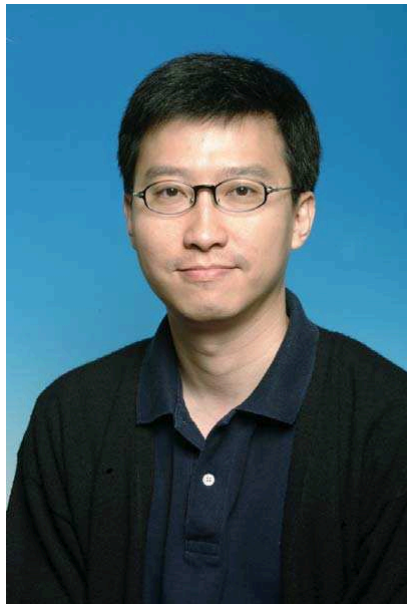
- [30] Turner SRE (1996) Resource Pooling in Stochastic Networks. Ph.D. Thesis, Statistical Laboratory, Christ's College, University of Cambridge
- [31] Turner SRE (1998) The effect of increasing routing choice on resource pooling. *Probability in the Engineering and Informational Sciences* 12:109–124
- [32] Vvedenskaya ND, Dobrushin RL, Karpelevich FI (1996) Queueing system with selection of the shortest of two queues: An asymptotic approach. *Probl Inf Transm* 32:20–34
- [33] Vvedenskaya ND, Suhov YM (1997) Dobrushin's mean-field approximation for a queue with dynamic routing. *Markov Processes and Related Fields* 3:493–526
- [34] Vvedenskaya ND, Suhov YM (2005) Dynamic routing queueing systems with vacations. *Information Processes, Electronic Scientific Journal. The Keldysh Institute of Applied Mathematics. The Institute for Information Transmission Problems*, vol 5, pp 74–86

Quan-Lin Li is Full Professor in School of Economics and Management Sciences, Yanshan University, Qinhuangdao, China. He received the Ph.D. degree in Institute of Applied Mathematics, Chinese Academy of Sciences, Beijing, China in 1998. He has published a book (*Constructive Computation in Stochastic Models with Applications: The RG-Factorizations*, Springer, 2010) and over 40 research papers in a variety of journals, such as, *Advances in Applied Probability*, *Queueing Systems*, *Stochastic Models*, *European Journal of Operational Research*, *Computer Networks*, *Performance Evaluation*, *Discrete Event Dynamic Systems*, *Computers & Operations Research*, *Computers & Mathematics with Applications*, *Annals of Operations Research*, and *International Journal of Production Economics*. His main research interests concern with Queueing Theory, Stochastic Models, Matrix-Analytic Methods, Manufacturing Systems, Computer Networks, Network Security, and Supply Chain Risk Management.

John C.S. Lui (M93-SM02-F10) was born in Hong Kong. He received the Ph.D. degree in computer science from the University of California, Los Angeles, 1992. He is currently a Professor with the Department of Computer Science and Engineering, The Chinese University of Hong Kong (CUHK), Hong Kong. He was the chairman of the Department from 2005 to 2011. His current research interests are in communication networks, network/system security (e.g., cloud security, mobile security, etc.), network economics, network sciences (e.g., online social networks, information spreading, etc.), cloud computing, large-scale distributed systems, and performance evaluation theory. Professor Lui is a Fellow of the Association for Computing Machinery (ACM), a Fellow of IEEE, a Croucher Senior Research Fellow, and an elected member of the IFIP WG 7.3. He serves on the Editorial Board of IEEE/ACM Transactions on Networking, IEEE Transactions on Computers, IEEE Transactions on Parallel and Distributed Systems, Journal of Performance Evaluation and International Journal of Network Security. He received various departmental teaching awards and the CUHK Vice-Chancellors Exemplary Teaching Award. He is also a co-recipient of the IFIP WG 7.3 Performance 2005 and IEEE/IFIP NOMS 2006 Best Student Paper Awards.



Quan-Lin Li



John C.S. Lui

Figure 8: